

# PERSONNEL SELECTION

*Test and Measurement Techniques*

ROBERT L. THORNDIKE

# *PERSONNEL SELECTION*

---

*Test and Measurement Techniques*

# PERSONNEL SELECTION

---

*Test and Measurement  
Techniques*

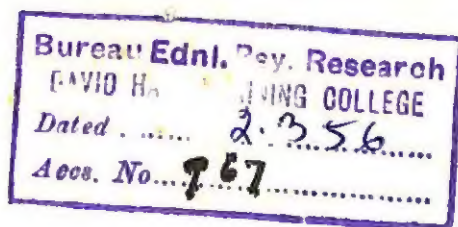
by

ROBERT L. THORNDIKE

Professor of Education  
Teachers College  
Columbia University

---

NEW YORK • JOHN WILEY & SONS, INC.  
LONDON • CHAPMAN & HALL, LTD. 1949



658.311  
T110

FOURTH PRINTING, DECEMBER, 1954

Copyright, 1949, by John Wiley & Sons, Inc.

All rights reserved. This book or any part thereof must not be reproduced in any form without the written permission of the publisher.

Printed in the United States of America



## Preface

This volume is an outgrowth of my four years of service in the Army Air Force during World War II. During that time I worked on the coordination of research on air-crew selection in the Aviation Psychology Program. A report specifically on the problems encountered in that program and the techniques adopted for dealing with them has already been published as *Aviation Psychology Program Research Report #3: Research Problems and Techniques*.<sup>1</sup> The present volume is based on that report but undertakes to formulate the problems and procedures in much more general terms which will be applicable to civilian as well as military personnel problems. In addition, a number of gaps in the earlier volume have been filled in, so as to present a more balanced treatment of testing problems in personnel selection.

I have used essentially the material of this book for the last three years for a course in the theory of measurement. The measurement problems in education and in personnel psychology have much in common, and most of the topics discussed in the following chapters refer to both.

This book is made up of two rather distinct parts. The first eight chapters deal with the technical problems involved in developing a personnel testing program and in appraising its effectiveness. The last three deal with administrative problems of maintaining an efficient, smooth-running program with good public acceptance. I believe that it is important for the worker in personnel testing to be aware of the problems in both areas. Both must be taken into account in planning an effective working organization.

Throughout the book, a number of problems have been raised for which I do not have an adequate answer. This is probably somewhat unsatisfying for the reader, particularly for the less advanced student of the field. But awareness of ignorance is

<sup>1</sup> U. S. Government Printing Office, 1947.

surely the beginning of wisdom, and perhaps pointing up some of our needs may stimulate a creative worker to help meet them. In particular, we have at the present time no very clear logic for the process of *classification*, as distinct from that of *selection*. If the preliminary suggestions in that direction are a stimulus to further productive thinking on the problem, the discussion will have been abundantly justified.

It has been difficult at times to decide just how to treat the statistical issues that have arisen in the course of this volume. This is not a text in statistics. On the other hand, some of the most troublesome problems relating to personnel selection are essentially statistical problems. In general, a background corresponding to about a year of work in statistics is assumed in the reader, and no instructions are given in the calculation of simple statistics. Where more advanced statistical procedures are needed, they are outlined, but no proofs or derivations are given. A number of chapters require almost no statistical background. The result is probably an effect of some unevenness as far as the statistical load is concerned.

The debts to colleagues who worked in the Aviation Psychology Program with me are numerous and cannot be acknowledged in full. The stimulation of many problems and of many able minds working cooperatively on them was a richly rewarding experience. In a very true sense, this book is a co-operative product though it appears with a single author named. I should like to mention especially Dr. J. C. Flanagan, Dr. A. P. Horst, and Dr. J. P. Guilford, under each of whom I worked at some time during the war and to each of whom I am indebted for many ideas which appear throughout the book.

ROBERT L. THORNDIKE

April 1949

# Contents

## CHAPTER

<b>1. INTRODUCTION</b>	<b>1</b>
Steps in the Development of Test Procedures	3
Problems in Operation of a Testing Program	8
<b>2. JOB ANALYSIS</b>	<b>12</b>
Job Description	14
Job Analysis	16
<b>3. TEST SELECTION AND INVENTION</b>	<b>32</b>
Use of Existing versus Construction of New Testing Instruments	32
Sources of Information about Existing Tests	33
Approaches to the Assembly of a Test Battery	36
Media of Testing	39
Rating and Observation as Selection Instruments	44
Steps in the Construction and Analysis of a New Test	49
Preparation of Objective Test Items	60
<b>4. THE ESTIMATION OF RELIABILITY</b>	<b>68</b>
Logical Considerations in Evaluating Reliability	69
Procedures for Estimating Reliability	78
Factors Influencing the Reliability of a Test	96
Interpretation of Estimates of Reliability	102
Need for Data on Reliability	104
Special Problems in Reliability Determination	111
<b>5. THE ESTIMATION OF TEST VALIDITY: CRITERIA OF PROFICIENCY</b>	<b>119</b>
General Problems in Connection with Criteria	120
Evaluation of Criterion Measures	124
Types of Criterion Measures	132
Summary Evaluations	149
Concluding Statement	159
<b>6. THE ESTIMATION OF TEST VALIDITY: STATISTICS OF VALIDITY</b>	<b>160</b>
Computation of Validity Indices	160
Taking Account of Restriction of Range	169
Curtailment with a Dichotomous Criterion	177
Non-Linear Relationship	180



## CHAPTER

<b>7. COMBINING TESTS INTO A BATTERY</b>	185
Factors Determining Multiple Correlation	190
Alternate Procedures for Using Test Data in Selection of Personnel	193
Length of a Test Battery	201
Addition of Tests to an Existing Battery	205
Combining Data from Partial Criteria	210
Role of Non-Statistical Factors in Weighting Tests	212
Three Major Types of Testing Programs	213
<b>8. THE ANALYSIS AND SELECTION OF TEST ITEMS</b>	227
Factors in Item Evaluation	227
Procedures for Calculating Item Indices	233
Use of Item Validities	243
Use of Internal Consistency Data	252
<b>9. THE ADMINISTRATION OF A TESTING PROGRAM</b>	257
Administrative Problems in the Conduct of Testing	259
Administrative Procedures in Test Scoring and Weighting	272
Organization of Reports and Records	286
Concluding Statement	293
<b>10. ADMINISTRATIVE PROBLEMS IN USING THE RESULTS OF AN APTITUDE TESTING PROGRAM</b>	294
Types of Necessary Administrative Decision	294
Factors Influencing the Administrative Pattern	296
Daily Quota versus Predicted Yield as Administrative Patterns	299
Inclusion of Non-Test Variables in the Prediction of Job Success	305
The Form of Test Score to be Reported	307
Function of the Personnel Psychologist in the Administrative Use of Test Scores	310
<b>11. THE PERSONNEL SELECTION PROGRAM AND THE PUBLIC</b>	312
Showing the Relationship of Test Scores to Job Success	314
Indices of Practical Effectiveness of Selection Procedures	323
Putting the Facts Together	330
Promotion of New Personnel Projects	331
The Personnel Psychologist and the Employee	333
<b>APPENDIX A. Solution of Normal Equations in Order to Determine Regression Weights and Multiple Correlation</b>	335
<b>APPENDIX B. Table for Estimating Correlations, Based on Upper and Lower 27 Per Cent of Group</b>	345
<b>INDEX</b>	353

## *Introduction*

The past thirty years have seen the birth, childhood, and coming of age of objective aptitude and achievement tests as devices for personnel selection. Tests of this sort came into being at the time of World War I. They were nourished during the period between the wars for educational selection in colleges, for employee selection in various industries and public personnel agencies, and for intellectual exploration by psychologists interested in tests and measurements. They took on adult burdens of responsibility in World War II in all branches of the armed forces. Hardly a man of the 14,000,000 who served the nation during that period failed to have his career affected in some measure by selection and classification tests devised and administered by personnel psychologists. Certainly in quantity, and probably in quality and diversity, personnel testing reached an all-time high. The record of wartime performance suggests that a rich future lies before the psychologist in the use of tests for selection and classification of personnel for industry, civil service, and education. This volume is concerned with the problems of testing such personnel.

Much of the discussion and, particularly, many of the illustrations in this book are set in the military frame of reference. This is in part a reflection of the general preoccupation of personnel psychologists during the five war years with the problems of military personnel. In part it is a specific reflection of the experiences of the author in the Aviation Psychology Program of the Army Air Forces. However, this setting is incidental rather than central to the purposes and values of the book. Personnel selection in the military differed from selection in industry in scope and in specific types of job duties rather than in basic principles. The steps and principles for developing, testing, and

applying personnel selection procedures are much the same in the two cases.

The use of tests in personnel selection has two aspects, the research and developmental and the operational. A testing program must first be devised and tested out; it must then be organized effectively for day-to-day testing and the efficient use of the test results. The first part of this book, Chapters 2 through 8, deals with the several steps in devising and testing out a selection procedure. Chapters 9 through 11 are concerned with administrative aspects of a testing program.

The feature that distinguishes reputable work in personnel selection from that of the mass of self-styled "psychologists," "personnel experts," and other quacks is that the reputable worker in the field is continuously concerned with testing, verifying, and improving the adequacy of his procedures. He knows that he does *not* know all the answers, and he is ever on the alert to find out more and to improve his procedures. There is no easy road to scientific personnel selection. The road is long and tortuous and beset with many pitfalls. Chapters 2 through 8 attempt to provide a guidebook to show the way that must be traveled in developing a selection program and to point out some of the rocky spots and morasses which lie in wait for the unwary.

Even after appropriate procedures have been determined, testing on a large scale remains an impressive administrative undertaking. How shall we guarantee uniformity of testing conditions, ruling out the effect of the particular examiner and equipment? How shall scoring be organized for maximum efficiency? What checks shall be provided on the accuracy of testing and scoring procedures? In what form shall records be kept so as to be most useful for a continuing research program as well as for day-by-day operations? In what form shall test results be reported, and how shall they be combined with other items of information about the individual? What shall be the policy with regard to action based on test results? How shall the use of test results be adapted to the factor of supply and demand in the labor market? How shall the results from the use of tests be shown? How shall the effectiveness of a selection program be interpreted to the interested layman? These and



other operational problems constitute the subject matter of Chapters 9 through 11.

The remainder of this chapter will be devoted to a quick over-view of the steps involved in the development, evaluation, and operation of a personnel selection and classification program. Thus, it will sketch in broad outline the picture for which the following chapters will fill in the details.

## STEPS IN THE DEVELOPMENT OF TEST PROCEDURES

### *Analysis of the job*

Personnel testing is done with the aim of selecting certain individuals from among the applicants for a job<sup>1</sup> or of determining to which of two or more possible job categories a particular individual shall be assigned. Before tests can be devised, or any other procedures set up for such selection or classification, the investigator must know as much as possible about the job in question. What does the worker do? under what conditions? What are the intellectual demands of the job? the physical? the social? Does the worker supervise, or is he supervised? Does he deal with other people or with machines? Is there pressure of speed? of complexity? Is the task repetitive or varied? These are only some of the questions to which answers should be forthcoming before the investigator undertakes to build up a program of tests to select individuals for this job. Evidently, profitable hypotheses as to functions important for success and tests useful for measuring them must arise out of an intimate knowledge of the job itself. The foundation for any selection and classification program is a rich background of information about the job or jobs with which the program has to deal and a discriminating analysis and organization of that information. Procedures for job analysis are discussed in Chapter 2.

<sup>1</sup> "Job" is used here, and throughout the book, as referring to a category of employment in an organization, rather than a specific single position. That is, the job is "stenographer," and there may be 1000 girls who fall in the category, rather than "Mr. Jones' secretary," who can be only a single individual. Job may also refer to a program of training preparatory to employment, such as the training in medical school, engineering school, and the like.

### *Selection and invention of test procedures*

After the research worker has become familiar with a job and made an analysis of the qualities required by it, his next task is to select or devise tests for those qualities. In some cases, an existing test may appear satisfactory as a measure of a particular function. Thus, if verbal comprehension is judged important for the job, one of the standard vocabulary tests might be selected for trial in an experimental test battery. In other cases, existing tests may lack some characteristic which seems important for the job in hand, and it may be desirable to build a new test which is specifically tailored to the particular job or testing situation. For example, in building tests for use in the selection program of the AAF, there was a general tendency to build tests around aviation content. Judgment tests dealt with flying situations, reading comprehension selections had to do with navigation or aviation equipment, mechanical comprehension tests replaced a diagram of a truss bridge with a diagram of a truss roof for an airplane hangar. These adaptations probably represented in some measure changes in the function measured by the test; in considerable part they were concessions to a demand for "face validity."<sup>2</sup> In still other cases, there may be no test available that corresponds satisfactorily to a function which it seems important to test. The research worker is then truly put upon his mettle to originate a new pattern of test performance and to develop a crude test idea into a practical and reliable testing instrument. This constitutes the most exacting and at the same time the most interesting and rewarding phase of test development work.

In any case, whether the test is genuinely new or merely a new form or minor revision of an existing test, specifications for each printed test must be set, in terms of type, number, and difficulty level of test exercises; test items must be written, revised, and edited; and the items for each test must be assembled into a test booklet. Apparatus tests must be designed, apparatus

<sup>2</sup> "Face validity" means that quality in a test which makes it appear sensible for the use to which it is being put, both to the subjects who are being tested with it and to the laymen in positions of administrative power who have the authority to decide whether the testing program shall or shall not continue to receive support.



built, and instructions and testing procedures formulated. The problems of test selection and development are considered in Chapter 3.

### ***Preliminary tryout and refinement***

When a new test form has been prepared, some degree of preliminary tryout, revision, and refinement is usually needed. This preliminary work becomes particularly important when the form of the test is very novel so that there is little background for judging how subjects will react to the items, whether they will comprehend the instructions, and the like. In this case, the original crude procedures may often be tried out on available clerical and office personnel to locate gross misunderstandings and to determine in a general way the level of difficulty at which items should be pitched. As a more refined form is developed, more extensive trials are usually appropriate, with groups similar to those for which the test is ultimately intended.

In careful work, it is generally desirable to submit the pool of items that have been developed for a test to a rather complete *item analysis*, to locate items that are too easy or too difficult and items that fail to discriminate between the more and less able members of the group. This analysis permits the test to be pitched at the appropriate level of discrimination and minimizes the wastage of testing time through faultily constructed items. Techniques of item analysis are presented in Chapter 8. In addition to the critical analysis of the component items, certain statistics, particularly estimates of reliability, will be desired for the test as a whole. The problem of obtaining suitable reliability estimates is considered in detail in Chapter 4.

### ***Validation of test procedures***

Whenever a test is being tried for selection of personnel for some job specialty, it is most desirable that it be validated empirically. Experimental evidence is called for to show that the test is in fact effective in discriminating between those who are and those who are not successful in a particular job. Though it may be necessary under the press of an emergency to rely upon the professional judgment of the psychologist to establish the value of a test for personnel selection, this must be recognized as

a stop-gap. A concern with validation, with seeking to subject all procedures to empirical test, is the mark of the sincere research man in personnel work.

Three aspects of validation require scrutiny. These are the establishment of criterion measures of proficiency, the administrative problems of assembling an adequate store of test and criterion data, and the statistical problems of analyzing those data and determining appropriate indices of validity.

The key to effective research in personnel selection and classification is an adequate measure of proficiency on the job. Only when proficiency measures can be obtained for the individuals who have been tested is it possible to check the effectiveness of test and selection procedures. The tests to be used for selection of aircraft pilots can be determined only by relating test scores to some later index of skill in the actual job of piloting a plane. The appropriateness of tests for picking insurance salesmen can be verified only against some such later record as actual amount of insurance sold. It might seem that such records of proficiency or performance would be so generally available as to constitute no particular problem in personnel research. However, this is not true. It is the general experience of workers in personnel research that finding or gathering relevant, reliable, and administratively practical criterion measures is the most difficult single task which the personnel psychologist faces. Problems of obtaining criterion scores are discussed in Chapter 5.

The administering of tests for personnel research and the assembling of test and criterion records for subsequent analysis present a number of practical administrative problems. These are primarily problems of logistics, of obtaining the required number of experimental subjects, and of assembling criterion data upon them with a minimum of delay. After applicants for a job are tested, months or even years may pass before each man has had a chance to demonstrate his level of skill on the job. Furthermore, the assembly of enough cases to provide stable and dependable statistical results may entail quite a period of testing. This is true particularly when the flow of personnel into a job specialty is limited. However, even in such a testing program as that of the AAF in World War II, in which half a million men were tested in four years, the pinch of limited numbers was felt

at times. It is important, therefore, that the most efficient possible use be made of the subjects available for testing.

After test and criterion data have been assembled, the next problem is to extract the most useful index of the effectiveness of a test for predicting the criterion. The usual correlation indices will generally be found appropriate at this point. However, several special problems arise in their use. We sometimes have to deal with dichotomous criteria, criteria which present a twofold split into some such categories as "graduates" and "eliminees," and this calls for an adaptation of statistical procedures. More seriously, the individuals who are available as a criterion group are often in some degree restricted or selected on tests or other bases, so that they are not representative of the complete group of applicants to whom tests will be given and from among whom future selections must be made. When high standards of selection prevail, this factor can become quite a serious one and can produce a very real distortion of the validities of different tests. These and other statistical problems in test validation are covered in Chapter 6.

### *Combination of tests into a battery*

In almost any situation in which tests are being used for personnel selection and classification, there will be a number of tests that are candidates for the selection battery. The problem is to determine what efficiency of prediction can be obtained from the tests together in teams and combinations, and with what weights the tests should be combined to yield the maximum accuracy of prediction. These questions may be asked concerning the complete pool of tests for which data are available, or they may be asked with reference to a more limited group of tests. That is, if practical considerations limit the testing time available in a particular selection program so that only three or four tests are to be used, we will wish to know which three or four of the available tests should be chosen, and in what way they should be combined. The problems involved in selecting tests and combining them into a composite prediction are elaborated in Chapter 7.

The basic factors which must be taken into account in combining tests for personnel selection are the validities of the

separate tests and the correlations between them. These factors enter explicitly into the procedures of multiple correlation and multiple regression, which constitute an objective and analytical approach to the use of a battery of tests for selection. They are also implicit in selection by means of cutting scores or in the clinical use of test scores. Some attention must be given to these alternative procedures for using a battery of scores. In each case, the significance of correlation between separate test scores must be appreciated as the major factor limiting the gain in accuracy resulting from the combination of a number of tests.

The use of tests for classification of personnel, that is, for determining which one of a number of jobs each individual is best fitted to do, introduces certain new problems in statistical analysis and procedure. Though analytical solutions are, for the most part, not available for these problems, the problems are interesting and significant ones, and the general rational considerations will merit analysis even though statistical procedures cannot be indicated with certainty.

## PROBLEMS IN OPERATION OF A TESTING PROGRAM

As soon as research and development have progressed far enough to make it profitable to use the techniques, the personnel psychologist faces the practical and administrative problems involved in putting a continuing testing program into operation. The operational phase often overlaps the research phase. On the one hand, the pressure of time sometimes makes it desirable to initiate a testing program based on accumulated previous experience and the best available professional judgment before data become available from the current program of research testing. In some limited testing programs, especially where the job in question is a familiar one on which a substantial amount of previous work has already been done, it may even be defensible to short-circuit the research phase entirely and to base selection procedures upon the accumulated experience of previous investigators. The research and operational phases also often overlap since in a vital selection and classification program research and development is a continuing process. Research does not stop when operational testing is begun but continues through-

out the program, with the goal of continuous refinement and improvement of testing procedures.

### *Administration of the testing program*

In a large-scale testing program, of which the testing of military personnel during World War II represents an extreme example, administrative and operational problems bulk quite large. If the program involves a number of individuals located at a number of places, the integration of their activities to assure uniformity and efficiency of operation and effective application of test results calls for administrative routines which are complete, detailed, and carefully planned. Such a plan, which became familiarly known as an S.O.P. (Standing Operating Procedure), might include such points as the following:

1. The order and timing of tests, intermissions, and other activities.
2. The verbatim instructions for each test, and procedures for handling all common questions and emergencies.
3. The responsibilities and duties of each member of the testing team.
4. Routines for scoring tests, combining test scores, and checking procedures both of scoring and combining.
5. Precautions for maintaining the security of test materials.
6. Procedures for calibration and maintenance of any equipment used either in testing or test scoring.

In any continuing testing program, even of moderate scope, it is worth while to give careful thought to the organization of testing activities and to formulate plans and procedures explicitly. In Chapter 9 attention will be given to a number of features that enter into accurate testing and efficient conduct of the testing program.

The outcome of any testing program is a score or series of scores for each individual. The analytical considerations in arriving at the selection of tests and the weights to be attached to them are discussed in Chapters 6 and 7. The form in which the resulting score is to be expressed, however, involves largely practical and administrative considerations. What form of score will be most meaningful to those who have to interpret it? What form of score will be most convenient for record keeping? What



compromise between convenience and exactness will lead to the most effective use of test scores in further research and development? It will be desirable to consider the relative advantages of raw and of standard scores for expressing the results of single tests. With standard scores, a decision must be reached as to whether they are to be reported to one, two, or possibly three digits. Similar problems must be considered for the composite score resulting from the weighted combination of the several tests in the battery.

More important than questions concerning the form of the test score are questions concerning the manner in which the score shall be put to practical use. Any given personnel selection and classification situation represents a problem in supply and demand and a problem in flow. A certain number of individuals must be chosen for each of one or several job specialties from among a certain number of applicants. There is generally a more or less fixed deadline by which time the personnel quota must be filled. Test scores must be obtained and utilized in such a way that the various quotas can be met by the specified times. These administrative demands may vary greatly from one situation to another. In some cases, it may be necessary to make an almost day-to-day adjustment of selection to immediate personnel demands. In other cases, such as that of a college recruiting a student body, there may be a time lag of weeks or even months between the date of testing and the date on which an administrative decision with regard to a particular individual must be reached. In some cases, an individual is a candidate for only one particular type of job, and the test scores are necessary only to select the most promising of the available applicants. In others, each individual may be a candidate for many jobs, and in using the score results an effort must be made to allocate the available personnel so that the over-all aptitude in all job specialties is the maximum.

Frequently, data and considerations other than test scores are also to be taken into account in the final decision as to personnel selection or classification. Such factors as age, education, preferences as to type of employment, record and recommendations from previous employment, and ratings and evaluations by interviewers may be relevant considerations. A

decision must be reached whether these are to be incorporated with the tests into a single composite score, or whether they shall supplement the test results in some other way. Finally, some decision must be reached on the allocation of final responsibility for assignment and employment. Shall these responsibilities be consolidated with those of testing and evaluation, or shall the personnel psychologist make recommendations to some further agency which will have the responsibility for the final administrative decision on each man? Some of these problems are discussed in Chapter 10.

A final operational responsibility of the personnel psychologist is to interpret his program and the results which he is obtaining to the non-specialist personnel in positions of higher responsibility in the organization, upon whom he relies for financial and organizational support. It almost always happens that those who have ultimate control over a personnel testing program have not had training in test procedures. They do not have the background to interpret correlation coefficients, much less regression weights. The rather involved statistical procedures of test research are likely to appear to them as something of a mystical mumbo jumbo unless the results from that research can be formulated in simple and explicit form. It becomes the responsibility of the research worker, therefore, to devise some relatively simple numerical indices or forms of graphic representation which will make clear to individuals in top administrative echelons what the testing program is achieving. In every personnel research program there must be a promotional campaign. No research worker can afford to neglect this promotional function, or he may find that he no longer has the opportunity to carry on his research. Therefore, Chapter 11 is devoted to methods of presenting test results to a lay audience, in order that the research worker who believes he is doing a worth-while job may have the best chance to show those in charge what his work is accomplishing.

## *Job Analysis*

In any program of personnel selection for a certain number of job specialties, the first step, logically and to a certain extent chronologically, is an analysis of the jobs in question to determine the activities which are carried out in those jobs and the circumstances under which they are carried out. From this knowledge of the activities and conditions of the job, and from his general background of psychological knowledge, the personnel psychologist derives insights and hypotheses as to the psychological functions required for success on the job and as to test procedures which might be appropriate to measure them. Thus, before the psychologist can hope to develop effective tests for the selection of an aircraft armorer, a drill-press operator, or a Comptometer operator, he needs to know as much as possible about what the worker does in each of those job categories. This knowledge must be not only complete but precise and specific to the duties in the particular plant or organization. The description needs to indicate not only the product produced but the processes gone through by the worker in producing it.

Knowledge about the job is indispensable as a source of insight about functions required in the job and consequently as a basis for selection or invention of tests designed to measure those functions. It is also essential for selecting or developing criterion measures of proficiency on the job. The general topic of criteria is discussed in detail in Chapter 5. We need only say at this point that such criteria are indispensable for any research and developmental program. We cannot do research on the selection of pilots until we are able to tell which of the men we have selected become good pilots. We cannot test a program for the selection of salesmen until we have some measure of the success of each man in the type of selling for which he was selected.



To determine what indices of success are already available and to obtain some estimate of their probable value, or to devise new procedures for evaluating success, calls for intimate knowledge of the job under study and of the records and other administrative procedures pertaining to it. One can evaluate ground-school grades as a criterion of pilot proficiency only as far as one knows what is taught in ground school (and how it is graded) and what the relationship of the material is to the skills of the pilot. One can evaluate the practical usefulness of production records in a given factory as a criterion for personnel proficiency only as far as one knows what records are available, how accessible they are, how accurately they relate to the individual worker, how nearly equivalent different machines and operations are, and the like.

As far as this volume is concerned, we are primarily interested in job analysis as a source of hypotheses for selection tests and as a source of insight about criterion measures. However, we should be aware that job analyses serve a number of other purposes in industry. One of the most widespread purposes of job analysis is as a basis for establishing pay schedules. Each job is analyzed to determine the extent to which various factors such as complexity, skill, training, and danger appear in it, and rates of pay are established by a weighting of those factors. Again, job analysis may serve the needs of a training program by indicating the specific skills and skill families which must be developed. It may serve as a guide to channels of promotion and of transfer from one job to another. It may serve as a basis for locating and minimizing safety hazards. It may provide the information needed for simplifying and regrouping the operations involved in producing some article.

The different applications of job analysis results call for the gathering of rather different sorts of material and arranging it in different ways.<sup>1</sup> It sometimes happens that the job analyses which the psychologist is carrying out as a basis for the development and evaluation of selection procedures may also be utilized for one or more of these other purposes. Then the information

<sup>1</sup> For a discussion of the use of job analysis for these varied needs, see C. L. Shartle, *Occupational Information, Its Development and Application*, Prentice-Hall, New York, 1946.

which is gathered and the form in which it is gathered may have to be modified to take account of the additional uses to be made of it. For application in a selection program, however, the information which a job analysis must provide is of two types. In addition to routine identifying information which provides a designation of the job category in which we are interested, we need (1) an accurate characterization of the job, that is, a *job description*, and (2) an examination of worker characteristics in relation to the job, to which we shall here apply the term *job analysis*.

### JOB DESCRIPTION

A job description is primarily a description of what the worker does and the conditions under which he does it. However, it includes a number of component elements or aspects. It should include first a statement in detail of the actual activities which the worker carries out. This statement should be comprehensive, covering every significant feature of the work that the worker must do as part of the job. The crux of the matter is the term "significant feature," and it is at this point that the judgment of the individual analyst enters in. Is it significant that this particular job requires the individual to read written instructions? That depends on the nature of the instructions and the nature of the group from which the candidates are being recruited. If the instructions are simple and the group is well educated, a group of college graduates, for example, the reading of those written instructions may not be at all critical. They may be well within the capabilities of every applicant. On the other hand, if the instructions are fairly complex or if the employee group is being recruited from applicants with limited educational background, the necessity of reading instructions may well become a very critical aspect of that job. A feature of a job is significant whenever inability to perform that feature might conceivably be a cause of failure in the job in members of the group from among whom employees are being selected.

The description of what the worker does should not only be comprehensive; it should also be specific. It should describe the major activities of a job exactly and in detail. In describing the work of the navigator in military aircraft, for example, the state-

ment "maintains a continuous log of the position of the plane, using pilotage, dead reckoning, celestial, and radio navigation procedures" is so meager as to be almost completely useless. This statement needs to be expanded to several pages, at least, indicating what types of observations must be made, what types of instruments, equipment, tables, and charts must be used, what types of calculations must be carried out, how frequently each must be done and under what conditions of pressure for speed and precision and state of emotional tension.

The report of what the worker does should provide not only a comprehensive and detailed statement of activities but also an indication of the importance of each. The description of the job should be organized in such a way as to facilitate identification of those activities and characteristics that represent *critical requirements* of the job, as distinct from those that are relatively incidental features of it. The following three indications of importance in a particular aspect of a job may be recognized:

1. The proportion of his time which a worker devotes to an activity.
2. The seriousness of the outcome if he fails to perform a particular activity satisfactorily.
3. The likelihood of some workers' being unable to perform the particular activity satisfactorily.

A description of a job should indicate not only what the individual does but also the conditions under which he does it. Is the work almost entirely sedentary or is it active? Is the work done largely apart from other people or is it in a group setting? Are there special characteristics of the physical environment, such as noise, dirt, or extremes of temperature, which may affect different individuals in their effectiveness on the task? Are there noteworthy psychological conditions, that is, such factors as danger, pressure for speed, or necessity for continuity of intense effort and concentration? A job description should provide an accurate picture of significant factors in the physical, social, and psychological environment in which the work must be carried out.

One feature of many job descriptions is a statement and sometimes a description of the materials and equipment which the

employee in that job needs. If a milling machine operator must work with calipers and a micrometer, that should be indicated. If he must read blueprints, that should be stated. If a statistical clerk is expected to use a computing machine, that should be specified.

## JOB ANALYSIS

After the description of the job has been completed, the job analyst should then interpret the job in terms of needed worker attributes. What qualities are called for if the worker is to be successful? This statement of desired qualities is, of course, a set of hypotheses. It is a set of inferences drawn from the job description. At the same time it provides a series of proposals for test construction.

There are two major problems in translating a job description into a good job analysis. One is to have a sound set of categories in terms of which to describe qualities of behavior. The other is to show sagacity in identifying those categories in the job description. Little can be said here to aid the development of that sagacity. It is the product of the individual's native wit and of all his psychological training, formal and informal. Experience in job analysis may certainly be expected to make a substantial contribution to it. The matter of categories, however, we may well discuss somewhat further.

What qualities are needed in a set of categories used to describe the attributes needed in a job? The following are suggested:

1. The set of categories should be comprehensive. It should cover the complete range of traits or qualities with which we may expect to be concerned in any of the jobs analyzed.
2. The set of categories should be organized and systematic. If each job analysis is to be complete, it is desirable to have a comprehensive outline of traits of human behavior. Reference to this outline and systematic review of the categories in it provides a check to make sure that no important points have been omitted from the analysis.
3. As far as possible, the categories should be independent. An efficient description of the individual calls for non-overlapping

traits. If several categories are much alike, using them adds to the complexity of the description without providing any corresponding increase in its completeness.

4. The categories should be psychologically meaningful. They should "make sense" in terms of behavior as we know it. This requirement is sometimes in conflict with our previous demand for independence, since the categories of common speech often are correlated. The factor analyses of human behavior carried out by Thurstone and many others represent an effort to reconcile these desired qualities of independence and meaningfulness, and many suggestions for useful psychological categories will be found in the factorial studies.

5. The categories should be of such a nature as to suggest testing operations for their measurement. We cannot, of course, neglect categories for which no measurement procedure is readily available. An effort should be made, however, to keep the categories close to behavior, so that they will be readily translated into a set of operations for their measurement.

An outline of categories is given in Table I, to suggest the type of organization that the analysis of traits required in a job might take. This outline is rather brief and skeletonized and has no special claim to being *the* outline of categories for a job analysis. It does, however, suggest something of the scope which should be covered in a job analysis and will guide the worker in the development of a set of categories for his own needs.

TABLE I. OUTLINE OF JOB ANALYSIS CATEGORIES

- I. Physical requirements.
  - A. Strength—in general, and by specific muscle groups where these become critical.
  - B. Endurance—resistance to fatigue.
  - C. Speed.
  - D. Gross coordination—in general, and for critical muscle groups.
  - E. Fine coordination—for specific muscle groups.
  - F. Adaptability—fluency in learning new motor patterns.
- II. Sensory requirements—acuity of each of the special senses.
- III. Perceptual requirements.
  - A. Speed of perception—for each sense as required.
  - B. Accuracy of discrimination—for each sensory attribute as required.



TABLE I. OUTLINE OF JOB ANALYSIS CATEGORIES (Cont.)

- IV. Intellectual requirements.
  - A. Verbal comprehension.
  - B. Numerical facility.
  - C. Deductive and inductive reasoning.
  - D. Mechanical comprehension.
  - E. Spatial visualization.
- V. Academic skill requirements.
  - A. Accuracy in mechanics of expression.
  - B. Fluency in verbal expression.
  - C. Mathematical knowledge.
- VI. Social requirements.
  - A. Pleasingness of manner and appearance.
  - B. Understanding of behavior of others.
  - C. Tact and deftness in dealing with others.
- VII. Interest requirements.
  - A. Interest in people.
  - B. Interest in mechanical things.
  - C. Interest in abstract ideas.
  - D. Interest in adventure, excitement, change.
- VIII. Emotional requirements.
  - A. Ability to function under pressures of speed, complexity, danger, etc.
  - B. Stability and personal adjustment.

### *Procedures for collecting job analysis information*

Information about a job can be obtained from a number of sources. Each has certain advantages and certain limitations. Five sources to which the job analyst may turn will be discussed here, and a section will be devoted to a critical review of each. The sources are:

1. Previous studies of the job.
2. Analysis of documentary materials.
3. Interviews with and interrogations of personnel.
4. Direct experience by the job analyst.
5. Statistical analysis of test validities.

### **PREVIOUS STUDIES OF THE JOB**

Science is a cooperative enterprise, and each scientist builds upon the work of his predecessors. This is true in personnel research as elsewhere, and the first resource in studying a job is other studies of that or closely similar jobs. There are several

places to which the research worker should automatically turn in looking for material of this kind. One is the extensive series of job studies published by the United States Employment Service.<sup>2</sup>

A second source is the *Psychological Abstracts*, under the job title or category. Among the psychological journals, the *Journal of Applied Psychology* and *Occupations* are probably the most fruitful sources of job analysis material. For military job specialties, much information is available in the reports, both of general and of limited distribution, which were prepared by the various psychological groups working in and for the armed forces during the war.

The amount to be gained from the review of previous studies varies from almost nothing to a substantial amount, depending on the particular job specialty in question. For many jobs of a more or less specialized nature, little directly relevant material is likely to be available, whereas more standard occupations may have been studied and reported on quite extensively. Where material is available, it provides a natural introduction to the study of the job. It has as outstanding advantages the facts that (1) the material is already available in complete and organized form and (2) the organization was presumably produced by a psychologist, who brought professional training and a professional point of view to the task.

There are, however, several limitations to this approach. In the first place, the literature provides an analysis of *a* job but not necessarily of *the* job in which we are currently interested. Even though the names are the same, the jobs may be different. The term "engineer" covers everything from a man who tends furnaces to a man who designs and invents electronic devices. It is only as firsthand acquaintance with a particular job is developed that we can judge the degree of identity of that job with jobs that are analyzed in the literature of personnel research.

<sup>2</sup> A list of available materials has been published jointly by the War Manpower Commission and the U. S. Office of Education. The reference is: *Guide to Counseling Materials*, U. S. Office of Education, Federal Security Agency and Bureau of Training, War Manpower Commission, May 1945. Some of these materials, as well as certain others, are listed in C. L. Sharple, *Occupational Information, Its Development and Application*, Prentice-Hall, New York, 1946, pp. 81-94.

Along a similar line, we can expect verbal job descriptions to take on full meaning only when the reader has some basis in personal experience for interpreting the verbal symbols. Just as a child's concept of an elephant must remain quite incomplete if he has never seen an elephant, so a psychologist's comprehension of certain job elements must remain pale and unreal if he depends entirely on words for his understanding of that job. Pilot instructors speak of "sense of sustentation" as a factor in flying and of "flying by the seat of the pants." How much meaning do these expressions have to a person who has never flown, or at least been flown, in a small plane? How adequate a picture can mere words give of the sequence of demands and pressures which a bombardier faces in synchronizing his bombsight during the last crucial seconds of the bombing run?

Some amount and type of direct personal experience is indispensable if the report of the work of others is to be meaningful and if its applicability to the present job is to be evaluated. Reports of the work of others are valuable, therefore, as a supplement to but not as a complete substitute for direct contact with the job to be analyzed.

#### ANALYSIS OF DOCUMENTARY MATERIALS

Two main types of documentary material are encountered in job studies. One type is made up of the instructional and operational manuals which are necessary in the operation of any extensive program of training or operations. The other type consists of records of the performance, achievements, or deficiencies of particular individuals or groups. As these two types of documents are quite different, it will be necessary to examine each in turn.

General training or operating manuals are often the materials most readily available for study of a job by a newcomer to a field. Whenever an extensive training program is carried out one may expect to find ready at hand courses of study, syllabi, textbooks, examinations, and the other printed devices for aiding and controlling the program. Operating procedures are often standardized in technical manuals and in codes of approved procedure. These materials were particularly prevalent in the



military setting, but their counterpart will also be found quite generally in civilian education and other civilian affairs.

The availability and convenience of these materials are their chief advantages. They may usually be obtained readily, and once obtained they may be taken home and absorbed at leisure. The limitations of the materials are fairly obvious. They are, after all, purely verbal presentations removed from the realities of the actual task. Like reports of previous studies of the job, the meaning they can convey to the reader is limited by his background of actual contact with the materials and operations to which the documents refer. They deal typically with rather gross units of behavior to be learned and results to be achieved rather than with detailed reports of activities to be carried out or of conditions under which they are to be carried out. They are concerned with results to be attained rather than personal qualities important for attaining them. They provide, therefore, only an indirect and rather remote set of cues to the actual psychological functions for which tests are desired.

The second type of record merits somewhat more detailed consideration. In any program of training or operations we may expect to find a number of types of records of the performance of individuals and groups. Most of these records are of interest in aptitude test development only because of their possible use as criteria of proficiency. They are in the form of quantitative grades, ratings, production figures, and the like, rather than in the form of qualitative records. As potential criteria for validation of selection procedures they are of great interest to the personnel psychologist, and he must become intimately acquainted with the characteristics of each type of record. One may also find, however, that certain types of qualitative and descriptive records have been maintained which throw light on the nature of job success or of job failure and of the traits which enter into such success or failure.

In the AAF pilot training program, "grade slips" were made out for each man for each flight. These grade slips contained not only quantitative grades on various maneuvers and phases of flying but also comments on the nature of the student's deficiency in any maneuver in which he was judged to be deficient. The reports of deficiencies were analyzed by aviation psycholo-

gists and tabulated by categories to provide evidence on the relative importance of various types of difficulties in flying.

Another type of record from which clues were obtained on the difficulties encountered and deficiencies revealed in learning to fly is represented by the Elimination Board Proceedings, available for each cadet who was eliminated from flying training. Here the instructor's testimony on particular deficiencies was available for each cadet. Analysis of 1000 of these Board Proceedings brought out the recurring patterns in instructor comments and provided a basis for setting up categories with regard to reasons for elimination.

A further type of record found for a few squadrons, and of particular interest because of its direct relevancy to the task of the flier in combat operations, was the report of reasons for failure of combat missions. For certain of the groups in combat theaters, mission reports were assembled and analyzed to show what had gone wrong on each unsuccessful mission. These analyses pointed to certain recurring deficiencies of combat personnel and provided suggestions as to factors to be considered in either selection or training of combat personnel or both, if the efficiency of the combat team was to be improved.

Civilian training and operational programs may well provide similar qualitative records of success and failure and of the reasons therefor. Types of records concerning which inquiry might be made include the following:

1. Qualitative reports by instructors or supervisors specifying the strong and weak points of students or of supervised personnel.
2. Case records of counselors and guidance workers, particularly in an educational program.
3. Separation interviews given at the time of leaving employment.
4. Reports of accidents involving personal failure.

Other personnel records useful for job analysis may be found in the particular situation.

Materials of the sort described in the preceding paragraphs appear to get somewhat nearer the determination of the actual psychological functions involved in a job than do general statements of training or operational procedures. It was possible to classify most of the comments on pilot grade slips or in Elimina-

tion Board Proceedings into categories according to the psychological function involved. Thus, certain remarks were classified as indicating deficiencies of memory, others as exhibiting defective judgment, and the like.

In the evaluation of materials such as Elimination Board Proceedings as they were used in the AAF, certain features made the records seem quite promising. In the first place, each record summarized impressions based upon a good deal of quite intimate experience with the man in question, since it was the final summary evaluation based on all his training at a particular training station. The evaluation typically combined the judgments of several men, instructors, and check riders, who had had a varied and often extensive experience of instructing cadets and checking their performance. The seriousness of the use to which the evaluations were put may be urged as evidence that they were carefully and deliberately made. The statements were subject to rebuttal by the student at the hearing, so that judges were under some pressure to make criticisms appear reasonable and appropriate to him. In other words, the evaluations represented an appraisal of individual ability which was the basis for an important practical decision and were therefore probably carefully and conscientiously rendered. In evaluating this type of record in other situations, points to be noted would be (1) the opportunity of the judge to observe the behavior, (2) the background and training of the judge, and (3) the amount of practical significance which the judge knows will be attached to the report.

However, such materials as the Elimination Board Proceedings and pilot grade slips also present several difficulties, both as to the adequacy of the original records and as to the interpretation of the reports in categories useful for test construction. In the AAF under the pressure of the wartime training program the number of these reports became very great, and it was felt that they became rather perfunctory and stereotyped. Instructors who were concerned with evaluating students developed a stock set of remarks which they applied, somewhat uncritically it was felt, to new individuals as they came along. Of course these stock remarks themselves represented a type of distillation of the essence of the instructors' experience, and they may have had

a certain basic validity for that reason as descriptions of fundamental student difficulties. However, they probably also represented in part the accidental growth of a tradition. Any job analysis based on case records consisting of stock phrases represents a rather remote second-order abstraction of a set of categories, since these stock remarks have themselves been abstracted from the concrete realities of experience.

A second difficulty with the AAF records was the tendency to make a strong case in an official hearing as a result of which administrative action was contemplated. If a person was being recommended for elimination, it was only natural to make the report on that individual as clear-cut and decisive as possible. It was felt that stock criticisms were often included in the report not because they were a particularly apt characterization of the individual being considered but because they were part of the accepted pattern of reasons given in connection with eliminations. This type of stereotyping must be borne in mind in connection with the use of any administrative records.

A final source of difficulty, which characterized both the approach to job analysis through records which is presently under discussion and the approach through interview and interrogation which is the topic of the next section, was semantic difficulty, that is, difficulty with language and meanings. Terms were used with meanings that varied from one report to another and that were in some measure at variance with the meaning of the same term to the psychologist working on the problem.

This can be well illustrated by the term "judgment." "Poor judgment" was repeatedly offered as a reason for failure in flying training, either upon a single maneuver or for the whole course of training. Further investigation into just what was meant by "poor judgment" revealed that it meant quite different things at different times and to different persons. In one case it meant lack of common sense as shown by a decision to fly through a storm rather than returning to the starting field. In another case it meant a faulty choice when the student was required to make an immediate selection of an emergency landing field for a simulated forced landing. In yet another case it meant inaccurate perception of the speeds and distances of planes in a traffic pattern. In still other cases it meant variations of these



and other types of judgments, intellectual or perceptual, which the individual was called on to make. It can be seen that the bare report of failure because of poor judgment could be only moderately instructive to the psychologist who was searching for hypotheses in terms of which to construct tests.

Language is a sufficient source of confusion in communication between trained psychological personnel; it becomes even more confusing when the psychologist is relying on various types of military or industrial personnel who have not been chosen on the basis of verbal facility and have not been trained to be precise or analytical in their reports of human behavior. The inarticulateness of worker and supervisor is a common difficulty in any method of job study that relies on communication from personnel engaged in the job.

#### INTERVIEWS WITH AND INTERROGATIONS OF PERSONNEL

Whenever the personnel psychologist studies a job, one important source of information will be the individuals who are learning, engaged in, or supervising the performance of the job being studied. These individuals represent a resource which is universally available whenever there is a job to be studied. Whether through informal and casual personal contact or through systematic interrogation, the analyst will find it desirable to assemble the testimony of those who have already had experience with the job. There are three types of individuals that should usually be considered as subjects for interview: persons who are proficient and experienced in the job under study, persons who are being trained to pursue the job, and persons who have failed in the task.

The values in interviewing competent and experienced personnel are fairly obvious. These are the individuals who, by virtue of their experience, are well acquainted with the job in its many phases. If they have had experience in instructing, supervising, directing, and coordinating the work of others, they are particularly likely to have a broad understanding of the task. Having engaged in the more advanced and specialized aspects of the job, they know what is demanded at the higher levels of performance. Thus, one could get insight into the job of training airplane pilots from instructors and supervisors of instruction,

whereas information about combat flying was almost necessarily obtained from those who were currently or had recently been in a combat assignment, preferably as a squadron or group commander or operations officer. In an industrial situation, one would look for job insight to the experienced workmen, foremen, and shop supervisors.

Learners may profitably be interviewed to determine the exact nature of their difficulties and problems. Since they are currently experiencing the difficulties, their awareness of them is more immediate and often more exact than that of the experienced workman who now experiences those difficulties only vicariously through others or in the far reaches of memory. For example, a focus of difficulty in the early stages of learning to fly is making landings. One helpful procedure for learning more about the specific nature of landing difficulties was to interview a number of students who were just going through the stage of learning to make landings.

Persons who are actually failing to make progress in the learning task provide another resource for learning about the critical difficulties of the task. The very fact of failure indicates that these individuals found the task particularly difficult and suggests that these individuals are in a particularly advantageous position to give testimony as to those difficulties. Their limited ability serves in a sense as a magnifying glass to pick out and emphasize those critical aspects of the task which are glossed over by the more experienced and competent. Though probably less able to provide a picture of the finished skill, failing students may provide valuable insights into the pitfalls along the way. Their report may, of course, be distorted somewhat by a tendency to rationalize their own failures by assigning responsibility to others and by their limited insight into the cause of their own difficulties.

These interview procedures have a great advantage over analysis of recorded material in that they are more flexible and permit the interviewer to pursue in detail any avenue that appears novel or promising. They make it possible to explore any leads suggested either by the interviewer or by the person interviewed. Again, if the individuals to be interviewed are selected with discrimination, it is possible to obtain from them an

otherwise unavailable richness and breadth of experience in the job being studied. Certainly the psychologist, who must be both a technical specialist and a student of many jobs, cannot expect to develop personally the degree of expertness in any one job which is attained by the specialists in that particular occupation. The real understanding of the job must come from the experts in it.

The effectiveness of interview procedures is very dependent on the qualities of the individuals interviewed. Many interviewees prove rather unrewarding sources. Articulateness and accuracy of expression together with an analytical approach to his job seem to be the important qualifications of an interviewee. In any interview, the psychologist is at least one step removed from the actual situation and can experience it only as it is reported to him by the person interviewed. The difficulty of communication is aggravated by the fact that in many jobs the specialists are not highly articulate about their own experiences. They are doers rather than tellers. Furthermore, their backgrounds may be quite different from that of the psychologist, so that there are differences in the referents and meanings of common terms.

In addition to the difficulties of getting from the job specialist effective expression in understandable terms, there are limitations in the analytical insight of the specialist. In a sense, interview procedures substitute the worker's untrained analysis of the functions which enter into his job for the trained analysis of the professional psychologist. Whether he is successful or unsuccessful, there may be reason to question the amount of insight which an individual has into the reasons for his success or failure unless he has had some special background or motivation to develop that insight. Responsibility for the training, supervision, and evaluation of others would, in general, appear to be good background experience for the development of such insights.

#### DIRECT EXPERIENCE BY THE JOB ANALYST

The fundamental resource of job analysis should always be firsthand experience of the job by personnel psychologists. Some degree of direct experience is almost a prerequisite for effective use of the secondary sources discussed in the previous paragraphs.

Reports of the experiences of others have meaning only in so far as the psychologist has some background of personal experience in terms of which to interpret these reports. Furthermore, the psychologist's professional training should make him the most apt individual at translating the job's demands into psychologically meaningful categories and into categories that may readily be transferred into testing operations. Experience by the psychologist may be of two main types, observation of the job and participation in the job. Each of these forms of experience must be considered for what it has to offer.

Observation of the job usually takes an early and a prominent place in any program of job study. The job analyst may sit beside the pilot and see what he does when taking off, flying, and landing the plane. He may ride with the bombardier and watch the sequence and tempo of his manipulations on a bombing run. He may stand beside the drill-press operator and follow the course of his operations of the machine. More will be learned if the worker explains step by step what he is doing as the job analyst watches. The understanding of a job which develops from such observation is probably a superficial one at best, but observation of the worker, inspection of his tools and instruments, and examination of his product provide a familiarization which enables the analyst to continue his study of the job through both direct experiencing and the indirect sources already discussed.

To obtain a fuller personal understanding of any job, the analyst needs to participate in it to some degree, as well as to observe it. He needs to fly the plane a bit, drop a bomb or two, try to operate the machine, or assemble a few of the gadgets. If time were of no concern, it would be desirable for him to take enough of the training so that he could experience at first hand the difficulties of mastering that job and developing the skills required of the competent operator. However, time is always a critical factor. The personnel psychologist is usually under pressure to develop selection procedures for a number of jobs, and, for each of these, job analysis represents only the preliminary step to the subsequent development and validation of test procedures. The amount of actual participation in the job by the



personnel psychologist is necessarily limited by such practical considerations.

The question of how far psychologists should go in mastering a particular specialty for which they wish to develop measures of aptitude or achievement raises a more general issue. The question is whether psychologists, in addition to acquiring a background of experience and highly technical training in appropriate psychological techniques, should also be expected to master the specialties to which they apply those techniques. It may be questioned whether such a philosophy of dual or multiple specialized training will be an efficient utilization of time and effort. The alternative is to draw heavily upon the background and experience of specialists in the various job categories. The experience of the psychologist serves then in large measure to provide a background which permits him to interpret and use the intensive training of the specialist.

From the practical administrative point of view, there seem to be very real advantages in enlisting the participation of operating specialists in developing the program of personnel selection and evaluation of their specialty. Interest in and support for the program will appear more spontaneously if responsible operating personnel have participated in its development. If the chief pilots or superintendents of flying have helped to develop the program for pilot selection, if the directors of maintenance have participated in developing the program for selection and evaluation of mechanics, or if the regional sales managers have had a hand in working out the problem of selecting salesmen, better administrative support for the program is almost guaranteed. Both from the point of view of necessary time limitations on the one hand and of organizational support on the other, therefore, there seems to be much to be said for leaving specialization to the specialist. The psychologist must have enough experience with the job to enable him to ask the specialist the right questions and to interpret the answers which he gets, but he should probably not try to become a specialist in each job that he studies.

#### STATISTICAL ANALYSIS OF TEST VALIDITIES

Some improvement of the research worker's insights concerning the important factors in a job may be obtained from an examina-

tion of the tests which are found empirically to have validity for that job. Although validity data become available too late in the cycle of test development to be of initial use in providing an understanding of job requirements, they provide a valuable objective check upon the initial hypotheses as to those requirements. As validity coefficients for different types of tests become available in considerable number, together with the test intercorrelations, a good deal of insight into the factors related to the criterion may be obtained from an examination of the correlational data. This insight may be further refined by the techniques of factor analysis,<sup>3</sup> including both test and criterion variables in the analysis. Study of the AAF classification test battery together with the available criterion measures of success in pilot, navigator, and bombardier training indicated, for example, that factors identified as "mechanical," "space relations," and "aviation interest" had the highest validity for the pilot, whereas "verbal," "numerical," and "reasoning" had substantially zero validity. In the case of the navigation criterion, high validities were found for the "numerical," "space relations," "science education," and "reasoning" factors, whereas "coordination," "aviation interest," and "visualization" had near zero validity. The designations of the factors are, of course, quite tentative, since the nature of each factor must be inferred from the tests in which the factor is found.

Analysis of validity data is valuable in clarifying aspects of the job performance which are already included in some degree in the existing tests. This may contribute to the improvement of a selection test battery by indicating factors for which improved and purified tests are needed. Again, tests may be found valid in combinations or for reasons that were not anticipated when the job was originally analyzed and the tests constructed, and thus understanding of the job may be extended. However, insights from an analysis of validity data are limited to the factors

<sup>3</sup> Factor analysis is a technique for analyzing the pattern of relationships among a set of variables, as shown by the intercorrelations of the variables, into a number of independent components or factors. For an introductory discussion of techniques see J. P. Guilford, *Psychometric Methods*, McGraw-Hill Book Co., New York, 1936.

that were in some measure included in those tests that were developed on the basis of the original job analysis. Within the scope of the original battery of tests, analysis of test validities and intercorrelations serves to check and refine the original job analysis, but these statistics do not provide a basis for extending the job analysis to new and virgin fields.

## Test Selection and Invention

After a study has been made of the job for which selection is to be carried out, leading to a description of the job functions and to hypotheses as to traits for which tests are needed, the next step is to assemble a battery of tests for experimental tryout and validation.

### USE OF EXISTING VERSUS CONSTRUCTION OF NEW TESTING INSTRUMENTS

It may be possible to select some of the tests for a battery from the existing pool of testing instruments; it usually is necessary to invent and construct some of the tests specifically for the particular selection project. Whether to use an existing test in any given case or to construct a new one depends on a number of different factors. Relevant considerations are the appropriateness of existing tests, the degree to which speed is an urgent consideration, the importance of maintaining the secrecy or "security" of the tests, the number of individuals to be tested, and the facilities available for new test construction.

The use of an existing testing instrument has a number of advantages. Perhaps the most obvious is availability. The test is available in printed form for immediate use. If the population to be tested is rather small it will also probably represent a decided economy to buy copies of an existing test rather than invest the time necessary for the construction of a satisfactory new test form. An advantage in some cases is that most existing standardized tests provide norms for various types of groups, so that there are standards with which to compare the present population. The chief disadvantage of a ready-made test is that it may not be exactly adapted to the needs of the present testing

situation. Thus, in the AAF air-crew selection program, rather than use an existing mechanical comprehension test, a new test was developed around planes and aviation situations. This test had much more face validity (apparent relationship to air-crew duties), and the specific content probably also contributed some additional component of actual validity. For flight engineer selection the general mechanical test seemed less appropriate than one dealing specifically with electricity and electrical systems, so a test of that type was prepared. For a large-scale testing program it is often worth while to develop a test specifically tailored to the needs of the unique situation.

Another factor which may prove a limitation in the use of commercially available tests in a personnel selection program is the availability of these tests to the public through various sources. It is more possible to maintain the security of test materials if the test has been specifically prepared for and is completely controlled by the present testing program. In civil service testing, where a good deal is at stake for the examinees, this consideration of security of materials becomes all-important. The administrative problems in connection with this problem of security will be considered further in Chapter 9.

There may be certain functions that are deemed important for the job under study for which no one of the pool of existing tests appears to provide a satisfactory measure. In that case it is, of course, necessary to construct a new testing instrument. These relatively novel types of test functions place the severest demands on the test constructor and at the same time provide his richest intellectual rewards.

## SOURCES OF INFORMATION ABOUT EXISTING TESTS

Effective use of the stock of ready-made tests depends on adequate sources of information about them. In recent years this information has been gradually better organized for the convenience of the test user. There are a number of sources of information with which any serious user of tests should be acquainted. First of all, mention may be made of the standard textbooks on psychological and educational measurement. Almost all these provide descriptions of and references to a number

of the standard tests of different aspects of aptitude and achievement. It is difficult to select specific titles for mention from among the many candidates, but the following list of annotated titles should help to guide the newcomer in the field to useful material.

G. M. Whipple, *Manual of Mental and Physical Tests*, Warwick and York, Baltimore, Vol. 1, 1914; Vol. 2, 1915.

An early book on testing which provides a very useful compilation of tests of simple psychological functions. Describes a number of procedures, in most cases individual laboratory procedures, for testing physical and motor capacity, sensory capacity, attention and perception, association, suggestibility, imagination, and the like.

P. M. Symonds, *Diagnosing Personality and Conduct*, D. Appleton-Century Co., New York, 1931.

Provides a full discussion of observation, rating, questionnaire, and test procedures for evaluating personality traits. Tests are described, and evaluative material with regard to each test and type of test is presented.

H. E. Garrett and M. R. Schneck, *Psychological Tests, Methods, and Results*, Harper & Bros., New York, 1933.

Describes a variety of tests, including tests of simple functions but emphasizing tests of complex abilities and traits. Provides fairly extended discussion of the problems and results of testing in different fields.

E. B. Greene, *Measurements of Human Behavior*, Odyssey Press, New York, 1941.

Describes, illustrates, and reviews typical tests of all types of psychological material. Noteworthy for the large number of reproductions of actual samples of test material.

H. A. Greene, A. N. Jorgensen, and J. R. Gerberich, *Measurement and Evaluation in the Secondary School*, Longmans, Green and Co., New York, 1943.

Test materials described are primarily those likely to prove useful in a school situation. However, many of the tests will also be of interest for use in non-school situations.

D. G. Paterson, G. G. Schneider, and E. G. Williamson, *Student Guidance Techniques*, McGraw-Hill Book Co., New York, 1938.

The book consists largely of brief descriptions and evaluations of tests. Though these are chosen with a view to their usefulness in the guidance of high-school pupils, the material on scholastic aptitude tests, vocational achievement tests, personality tests and questionnaires, and special aptitude tests will be of interest also for the research worker in personnel selection.



W. V. Bingham, *Aptitudes and Aptitude Testing*, Harper & Bros., New York, 1937.

The body of the text is devoted to a discussion of different types of job categories. An appendix is devoted to the description and evaluation of a number of specific tests. These are chosen for their usefulness with adults in a counseling or employment situation.

Herbert Moore, *Experience with Employment Tests*, Studies in Personnel Policy No. 32, National Industrial Conference Board, New York, 1941.

This is not a textbook but a report prepared specifically for those working on employment problems in industry. It describes a number of tests and evaluates them in terms of the success which has attended their use in industrial personnel work.

The most comprehensive bibliography on testing has been prepared by Hildreth.<sup>1</sup> This bibliography, which is organized by the type of function tested, lists all the tests which have been published or concerning which published material is available. No annotation or other comment is given with the references, however, so that the bibliography serves purely as a library tool in locating the basic sources of information about any one test or about tests of a particular function. For that purpose it is outstandingly useful. A somewhat less complete and less current bibliography, but one supplying an annotation about each test, has been prepared by Wang.<sup>2</sup>

Further information about published tests, together with critical reviews of them, may be obtained from the *Mental Measurements Yearbooks*.<sup>3</sup> The yearbooks of most importance to the test user are those of 1938 and 1940. Publication was discontinued

<sup>1</sup> G. Hildreth, *A Bibliography of Mental Tests and Rating Scales*, The Psychological Corporation, New York, 1939; *A Bibliography of Mental Tests and Rating Scales, 1945 Supplement*, The Psychological Corporation, New York, 1946.

<sup>2</sup> C. K. A. Wang, *An Annotated Bibliography of Mental Tests and Scales*, Catholic University Press, Peiping, China, 1939.

<sup>3</sup> O. K. Buros, *Educational, Psychological and Personality Tests of 1933, 1934, and 1935*, Studies in Education No. 9, Rutgers University, New Brunswick, N. J., July 1936; *Psychological and Personality Tests of 1936*, Studies in Education No. 11, Rutgers University, New Brunswick, N. J., August 1937; *The 1938 Mental Measurements Yearbook of the School of Education*, Rutgers University Press, New Brunswick, N. J., 1938; *The 1940 Mental Measurements Yearbook*, The Mental Measurements Yearbooks, Highland Park, N. J., 1941.

during the war but has been resumed, and it may be assumed that future volumes will also be fundamental sources of information for the test user. These yearbooks have been concerned primarily with reviewing recently published tests, though they have attempted to provide some coverage of earlier material. As a result, the tests covered in the yearbooks do not exhaust the whole list of existing tests in a given field. Most tests are critically reviewed by one or more reviewers, so that the reader obtains some description and evaluation of each instrument.

One source both for information about and supply of existing tests is the Psychological Corporation.<sup>4</sup> This organization, founded in 1921 by James McKeen Cattell as a pioneer organization to market psychological services, includes among its activities the publication of psychological tests and the distribution of a wide range of tests put out by other publishers. To a legitimate, bona fide professional user of tests the staff of the testing division will supply not only the tests but also some advice and information about them.

As a final source of information about existing tests, reference must of course be made to the test publishers. In addition to the Psychological Corporation there are many concerns which publish educational and psychological tests. A list of most of these can be found in *The 1940 Mental Measurements Yearbook* referred to in an earlier paragraph. The catalogues of these publishers indicate the items available from each. By obtaining a sample set, together with the accompanying test manual, the user can determine for himself the exact nature of the test and estimate whether the test will prove satisfactory for his purposes.

## APPROACHES TO THE ASSEMBLY OF A TEST BATTERY

In a thorough program of research directed toward the development of a testing battery to predict success in one or several jobs with maximum accuracy, some systematic approach is needed to develop tests that cover as many as possible of the functions disclosed by job analysis. A haphazard attack on the problem is likely to result in overconcentration of test develop-

<sup>4</sup> The address of the Psychological Corporation is 522 Fifth Avenue, New York, N. Y.

ment in certain areas and neglect of other areas of importance. There are two ways in which this systematic approach may be oriented. It may be focused on the job or on the individual. The former we shall label the *job approach* and the latter the *trait approach*. These two approaches use the results of job analysis and go about test construction in quite different ways. We must examine each to see what its characteristics and potentialities are.

Let us first examine the *job approach* to test development. In this approach, the research worker starts primarily from the job and the duties which the individual must perform in it. He tries to build tests that reproduce some feature of the job. Since in flying the pilot must use a stick and rudder pedals and coordinate the movements of the two, the developer of tests for pilots is likely to construct an apparatus that requires the subject being tested to use a stick and rudder bar and to make coordinated movements with them in response to some type of cue stimulus. Since the navigator must use various types of rather complex tables, a test of table reading is prepared. A test for street-car motormen may require reaction with one of several handles to cues from a motion picture of a street scene. In these examples the functions to be tested have been seen primarily in terms of the characteristics of the job and job duties rather than in terms of any fundamental pattern of human abilities. This approach yields a job sample type of test in which the tester endeavors to reproduce in a miniature situation all the complex conditions of the job itself rather than tests of relatively simple and general traits of behavior. In determining whether he has achieved complete coverage of the job, the investigator will say to himself, "This worker must do X on the job. Do I have a test in which he must do X or something very like it? In which test have I required the examinee to carry out a task similar to job function Y? I must put into one of my tests a task as much like Z as I can."

The *trait approach* stands in sharp contrast to the job approach in that the test development is based on the general qualities of the individual rather than on the characteristics of a specific job. The initial effort is to identify a set of fundamental categories or traits of human behavior, preferably non-overlapping traits in

the sense that each trait will have zero correlation with each of the others, and then to develop tests of those basic traits. Initially, hypotheses concerning human trait structure emerge from the general psychological study and analysis of human behavior. These hypotheses become refined by a study of the obtained pattern of relationships among tests designed to measure the hypothesized traits. This approach to test development yields tests measuring such functions as numerical facility, verbal fluency, and perceptual speed. They are tests of generalized functions, not specifically related to any one job. Research on personnel selection for a particular job involves trying out tests of all those traits that seem to have some relationship to success on that job. The job analysis is studied with a view to selecting the promising trait categories. Then tests of these categories are selected from the existing pool of tests or developed for the specific program.

The job and trait approaches, in their pure forms, represent opposite extremes of a continuum rather than unrelated approaches. In practice, most test development research falls along some intermediate range of that continuum, emphasizing to some degree both the reproduction of the conditions of the job and the analysis of the basic traits of human behavior. In general, the *job approach*, in which each test is built to the specifications of the particular job specialty which it is to predict, is likely to yield tests with higher validity for a specified job. However, the single tests for a particular job will tend also to have higher intercorrelations, so that there will be less gain from adding on supplementary tests. High correlation among the single tests also makes their use for *differential* prediction of a number of job categories more difficult. The advantages of the two approaches are, in part at least, relative to the manner in which the test results are to be used. A more complete evaluation of them must be postponed to Chapter 7, in which the use of a battery of tests for simple selection is contrasted with its use for multiple selection and for classification.

Both the job and the trait approaches provide frameworks for a systematic and complete coverage of the abilities important for a job. An integrated program will represent some synthesis of these two. However, in any research program, especially

where the research is being carried out by a number of persons and possibly in a number of places, some test development is likely to take place outside the systematic program. Although a systematic framework is a necessity for the effective organization of ideas and plans for test development and the reduction of duplication of effort, individuals inevitably have ideas for tests independently of this framework. Arising from some special interest or experience of a particular investigator, an idea for a test emerges. Such an idea may be worth developing, even though it does not fit into the systematic over-all program of test development. Because of the limitations of human insight in planning an inclusive, comprehensive program, the encouragement of isolated ideas, which arise without regard to the total program of test construction, is a sound procedure for extending the scope of test development.

## MEDIA OF TESTING

When we speak of a testing program, we first think of printed tests, in which successive test items present problems to the subject who is required to work out or identify the correct answer. This is the most usual and universal type of test, but there are others. Test problems may be presented on a motion picture screen, making possible the presentation of certain types of tasks that cannot be presented in printed form. Again, it may be desired to test skills of manipulation or coordination, and for these purposes apparatus tests may be required. Still other assessment procedures may involve interview or rating procedures. These may be relatively informal and uncontrolled, or they may require a highly standardized form of interview and procedures for evaluating interview responses. The individual test of the Binet type is a familiar illustration of the highly standardized interview procedure. This section will be devoted to a discussion of the values of different media of testing; the next section will be devoted to a consideration of interview and rating procedures.

Printed group tests have always occupied and will continue to occupy a major role in most mass testing programs. These possess as their outstanding advantage the efficiency and economy



with which they can be administered and the relative ease with which objectivity can be achieved in administration and scoring. The matter of economy and efficiency is particularly important in large-scale testing programs, but the matter of objectivity maintains its importance no matter what the scale of the testing.

Of course, some types of ability do not lend themselves to assessment by printed tests. The range of traits measurable by printed test techniques is not sharply defined, however, and with sufficient ingenuity it may be possible to develop printed test techniques for assessing various aptitudes that had previously been considered susceptible only to other more laborious methods of assessment. In the AAF air-crew testing program this was illustrated by measurements of a "spatial relations" factor. Analyses of the test battery in use for personnel classification together with groups of research tests indicated that a large fraction of the validity for pilots of several apparatus tests could be attributed to a "spatial relations" factor. Subsequently, several printed tests were developed which emphasized the orientation of planes in the air and of the ground as seen from planes, and these depended on this same "spatial relations" factor. Much, though not all, of the valid variance of the apparatus tests could have been covered by these group tests. If group test procedures could have been developed to cover the remainder of the valid variance contributed by the apparatus tests, the testing procedures would have been appreciably simplified. As a general guiding principle of testing programs, printed tests should be used for all functions for which they are adequate. Furthermore, it is usually worth while to devote some part of the creative effort of a testing program to inventing printed test procedures for measuring traits of importance not previously measured by those techniques.<sup>5</sup>

In spite of all efforts to broaden the field covered by printed group tests, there will probably always remain certain functions for which printed tests are not adequate. Types of perception

<sup>5</sup> The development and application of a wide range of printed tests to a particular personnel selection program is illustrated by the work on air-crew selection in the Army Air Forces during World War II. The tests have been described in detail and provide a rich pool of suggestions for workers in other specific personnel situations. Tests and results from using them

in which motion of the stimulus is a necessary feature call for motion picture tests. Printed tests also seem poorly adapted for use in almost all areas in which speed or coordination of motor response are significant features. Individual apparatus tests then seem indicated. Again, with printed tests it is almost impossible to devise procedures for timing accurately the exposure of each successive stimulus or the rate at which the stimuli are presented to the subject. Either motion picture or individual apparatus techniques can be devised to deal with this type of situation. Finally, where accurate timing of single responses by the subject is required, individual apparatus testing seems to be almost a necessity. Depending on the type of job for which selection is made, it may be desirable to devote some fraction of the test development research to investigation of motion picture and apparatus tests.

Motion picture tests retain most of the advantages of group administration which characterize printed tests. In certain phases of standardization and objectivity they even surpass the printed test. Instructions recorded on a sound track and accompanied by illustrative exercises on the screen reduce the variation in that aspect of the testing situation almost to the vanishing point. These tests possess in addition certain unique advantages. The most obvious of these is the possibility of introducing movement in the stimulus field. It is thus possible to construct a variety of tests of perception of rate of movement, direction of movement, and the like, as well as tests that simulate the experiences of an individual in a moving vehicle. These have obvious relevance to research on the selection of gun-pointers, pilots, truck drivers, and so forth.

A second conspicuous advantage of the motion picture is the control that this medium permits in the time allowed both for presentation of the stimulus and for response by the subject. The number of frames allotted to a particular test item determines the former, and the amount of blank film introduced

are described in the following reports: J. P. Guilford, *Printed Tests*, AAF Aviation Psychology Program Research Report #5, U. S. Government Printing Office, 1947; F. B. Davis, *The AAF Qualifying Examination*, AAF Aviation Psychology Program Research Report #6, U. S. Government Printing Office, 1947.

between items fixes the latter. It is possible, therefore, to control both the time allotment for any single item and the amount of over-all speed pressure of the test. A further characteristic which may be of advantage is realism. By using motion pictures with sound, it is possible to present test situations that appear more like the real life situation than can be done with any type of printed material. This seems particularly advantageous in proficiency tests, in which every increase in resemblance between the test and the actual job situation contributes to relevance of the test as a measure of on-the-job performance.

The limitations of motion picture tests are primarily practical ones connected with test construction and use. A first obstacle is the very considerable amount of technical skill required to produce an effective film. Though some photography, especially for preliminary forms, may be done by the psychologist, for the production of a satisfactory final test form it often is necessary to rely on the technical skill of professional studios. This involves not only expense but also certain problems of coordination and cooperation between the test constructor, who knows what effect he is trying to achieve for his test, and the technician, who has the skills necessary to achieve it.

Other problems arise in connection with the actual conduct of testing. These have to do with lighting and seating. There must be enough general illumination so that subjects can see to mark their answer sheets, and yet not so much as to sacrifice sharpness of definition of the screen image. Seating must be arranged so that no individual is penalized or favored because of his angle of vision or distance from the screen. Experience in the testing program for selecting air crews for the AAF showed these factors to be less critical than had been anticipated. Over rather wide limits position seemed to be unimportant, except in certain tests which made obvious demands on visual acuity. By the same token, rather a wide range of illumination levels seemed to be equivalent as far as test score was concerned, including levels at which the answer sheet could readily be marked. However, these factors would have to be explored for each novel type of motion picture test.<sup>6</sup>

<sup>6</sup> The use of the motion picture as a medium for testing was explored quite extensively in the Aviation Psychology Program of the AAF. A full report of this experience is given in J. J. Gibson, *Motion Picture Testing*

Individual apparatus tests appear necessary whenever we are primarily interested in the motor aspect of the subjects' response. Whenever the *making* rather than the *selecting* of the response is the important consideration, some instrument is usually needed to record the speed, precision, or other relevant features of the subjects' response. Measures of this type are often required in selecting personnel for a job in which manipulative skills play a critical role. Apparatus tests may cover the whole gamut in complexity. Simple form and peg boards represent these tests at the simplest level. The most complex type of apparatus test is suggested by some of the synthetic training devices used in the armed forces. Devices such as the Link Trainer, which reproduces for the student pilot all the characteristics of flying a plane by instruments, or the Celestial Navi-trainer, in which the student navigator has a continuous view of the sky and stars just as they would appear to him in a plane flying over a given route at a specified time, are examples of tests that very nearly reproduce the actual job itself.

The use of apparatus tests in personnel selection in general and in personnel selection research in particular raises a number of problems. An obvious one is the time and equipment required for any large-scale testing program. The usual testing situation requires one copy of the apparatus and one examiner to test one subject. In the AAF air-crew selection program, a considerable increase in efficiency was obtained by grouping four copies of an apparatus together in a testing room, with a single control table on which all controlling and recording instruments for the four copies were centered. In this way, one examiner was able to test four subjects at the same time. Even so, to administer a battery of six apparatus tests to a hundred men a day called for five copies of each test (including a spare copy in reserve for breakdowns) and the services of eight to ten examiners. Such a program clearly represents a substantial investment and must be justified by real increments in validity resulting from use of the apparatus tests.

A second difficulty with apparatus tests is that of standardization and maintenance. This point will be elaborated in Chapter 9. It will suffice at this point to say that if the several copies of  
*and Research*, AAF Aviation Psychology Program Research Report #7,  
U. S. Government Printing Office, 1947.

an apparatus test are to yield equivalent scores their construction must be accurately engineered. If a single copy of a test is to continue to give equivalent scores it must receive effective maintenance and frequent calibration.<sup>7</sup>

When apparatus tests are used for research, one additional major difficulty is encountered. That is the difficulty of assembling test records fast enough so that validation data become available upon an adequate population within a reasonable time. When a new printed test has been developed for research purposes, it is administratively quite simple to include that test in the battery being given to all personnel applying for the job under study. When a new apparatus test is developed, however, it may be available in only one or two pilot copies. Particularly if the test is complex or expensive, it is unlikely that a number of copies of it will be developed for research use alone. It is then often possible to test no more than 100 or 150 subjects a week with the new test. Conditions often require the testing of several thousand applicants in order to get an adequate population for whom criterion data will become available in a particular job specialty. (The practical problems of scheduling validation testing will be discussed in more detail later in the chapter.) Thus, testing over a period of several months may be required for the accumulation of sufficient testing data. When to this the time is added that must elapse while criterion data mature and while copies of the apparatus are procured, it can be seen that a research program for the development and use of validated apparatus tests is likely to be a slow-moving affair.

### RATING AND OBSERVATION AS SELECTION INSTRUMENTS

Certain types of behavior do not readily yield a record on any type of test. These are particularly the situations involving the

<sup>7</sup> Apparatus tests were used extensively for mass testing in the Aviation Psychology Program of the AAF. A report of much of the research work on the development and validation of these tests and some of the problems of standardization and calibration of equipment is planned for publication by the U. S. Government Printing Office as AAF Aviation Psychology Program Research Report #4.



social interaction of two or more individuals. The reaction of a subject to the situation of being interviewed, the manner of the subject when placed in charge of two or three other individuals, the response of the subject to interference, heckling, or cross-examination, and other types of responses can probably be evaluated only or most readily by the ratings of those who observe the behavior in question. Again, for some types of jobs we are concerned with the impression which an applicant makes upon other people. In the selection of salesmen, for example, an important consideration, in addition to what a person *can do*, may be what he *appears to be* as he encounters the prospective customer. Where the factor to be evaluated is the impression made on another individual, we can probably best evaluate that factor by determining the impression made on another individual in a situation as similar as possible to that occurring on the job.

A second purpose for which interview procedures may seem the most suitable is to obtain information about the life and job history of the individual. Information about the type of life an individual has led, his home and family relations, his school experiences and his reaction to them, his social activities and interests, and his background of work experience are often very important in the total selection and classification situation. This is especially true when one is concerned with employing not novices but rather men with a substantial work history, which may bear some relationship to their present assignment. It may be desirable to give as much or more weight to what the individual has done in life in the past as to what he can do on a specific battery of tests at the present time. Thus the Army and Navy, in planning for classification of enlisted personnel, found it important to get both records of civilian training and work history and objective test records of present aptitudes.

It is possible to obtain information about life history with group test procedures. A biographical data blank may be prepared in multiple-choice form which presents a number of fixed-choice questions to the subject and requires him to select the option or options applicable to him. Such a blank was used to advantage during World War II by both the Army and the Navy for the selection of airplane pilots. This type of blank has the complete objectivity of scoring found in any good aptitude or

achievement test. It is, however, quite inflexible and is limited in scope by the questions asked and the response options provided.

A second approach to personal data is through a job application blank or personal data questionnaire, on which the applicant makes unrestricted responses to a number of questions about his background, education, and training. Everyone has the experience of filling out many such blanks as he pursues his education and his job career. The application blank sacrifices the objectivity of a fixed-choice blank, which can be scored by a predetermined key, but it permits more flexibility and variety of response by the applicant. A somewhat subjective and clinical evaluation must then be made of this material in order to judge how well the individual is qualified for the job for which he is applying.

In addition to or in place of the above methods, a personal interview may be a source of biographical information. The interview permits still greater flexibility. It is possible to adapt the questions to the individual and to concentrate on exploring those aspects of his background that seem most important. Interesting leads may be followed up in more detail, and responses that are not clear may be further clarified.

The interview may yield a record, usually in the form of a descriptive report, of certain aspects of the individual's background. It may also result in a definite impression on the part of the interviewer as to how suitable the candidate is with respect to the different qualities required for the job. That is, an interview may yield both a description and a rating or evaluation of the individual's background. This evaluation is at once the strength and the weakness of the interview procedure. Only part of what occurs in an interview can ever become a matter of record. The interviewer, however, is reacting to the complete experience, the complex interplay of statement, intonation, manner, and the like. He is, therefore, probably in the best position to synthesize and evaluate the material brought out in the interview period. His direct judgment is more likely to be correct than is the secondhand judgment of someone else applied only to the recorded material of the interview. On the other hand, the very complexity and subjectivity of the interview situation makes it possible that the interviewer will react to non-

essential features and will be unduly influenced by the fluency, presentableness, or earnestness of the applicant. Certainly, the responsibility placed on the interviewer is very heavy, and skill and training are needed if the clinical synthesis of interview material is to be more effective than the objective treatment of life history data which is possible in an objective biographical data blank.<sup>8</sup>

The subjectivity and bias inherent in rating procedures sharply limit the value of these methods, so that they are not to be favored for functions that can be evaluated by any type of objective test. However, it does seem that observational, interview, and rating procedures help to extend the scope of evaluation. Nevertheless, in planning a rating procedure in a personnel selection program, one should first be reasonably sure that no objective test procedure is available for that function. To cite an extreme illustration, it would be foolish indeed to try to evaluate the extent of an applicant's vocabulary on the basis of an interview and rating rather than on the basis of a test.

The second prerequisite in using rating procedures is that the qualities to be rated must be those that the rater has had or will have an opportunity to observe. On the basis of the usual interview situation, the interviewer could rate the applicant with some assurance on neatness of dress, on poise in the interview situation, or on pleasingness of voice and diction. It would not be reasonable to ask for a rating on integrity, industry, or ability to get along with fellow workers. These are aspects of behavior that could hardly appear within the limited framework of the interview situation.

The range of behaviors exhibited during an observation or interview situation can be increased by presenting the subject with various task situations. He may be called on to work with others or to supervise the activity of others. He may be placed in a situation in which the task he must do is frustrating, or in

<sup>8</sup> A discussion of techniques of interviewing and of using interview materials is beyond the scope of the present volume. A general discussion of interviewing is given in W. V. D. Bingham and B. V. Moore, *How to Interview*, Harper & Bros., New York, 1941. A specific technique of interviewing and using interview results in personnel selection is described in R. A. Fear and B. Jordan, *Employee Evaluation Manual for Interviewers*, The Psychological Corporation, New York, 1943.

which it is made so by the observation or heckling of spectators. He may be asked to play one or more imaginary roles. He may be subjected to interrogation approaching the third degree. Situations of these and other types were combined with observation and rating procedures by the Office of Strategic Services in the selection of agents for special types of duty during the war and were considered to have been effective, though statistical validation in this case encountered a variety of difficulties.<sup>9</sup>

In addition to limiting ratings to behaviors evident during the interview or observation situation, the interviewers or observers must know what they are looking for and must have a common understanding of the scale of ratings. That is, the observers must become acquainted with the rating instrument and procedure *prior to the period of observation or interview upon which the rating is to be based*. This means a consideration of the qualities to be rated, of the symptoms of these qualities to be looked for in the rating situation, and of specimens of behavior representing a particular degree of a particular trait. When a number of observers or interviewers are to make the ratings, as is usually the case, preliminary training of raters is particularly important. Subjectivity and individual idiosyncrasy in ratings will be large enough in any case, and every effort must be made to reduce it. One procedure is to hold preliminary and periodic conferences and discussions among those doing the rating in which the meanings of the several scales upon which ratings are being made and of the separate points upon those scales are reviewed.

In general, the simpler and more overt the behavior to be rated, the more nearly objective we may expect the ratings to be. We may anticipate more agreement on rating or checking something that the individual *did* than in evaluating what he appeared to *be*. As far as possible, therefore, ratings and evaluations should be expressed in terms of observable actions.

The usefulness of rating procedures in a selection program may be evaluated on two bases, effectiveness and practicality. Effectiveness is limited by those factors of subjectivity and in-

<sup>9</sup> A full description of the procedures developed in the OSS assessment program will be found in the assessment staff report, *Assessment of Men*, Rinehart & Co., New York, 1948.

dividual bias which produce low reliability in ratings, and which can be only partially compensated for by a program of rater training and review. The fundamental question of the validity for a particular job of those traits which can be rated in an interview or observation session is, of course, a question that must be determined empirically for each job specialty.

It is from the standpoint of practicality that rating methods may be most seriously questioned, especially in a large-scale personnel selection program. Providing personnel to carry out the necessary observations or interviews is usually an expensive undertaking. Observations are time-consuming, and if they are to have value they must be made by capable and trained personnel. Therefore there must usually be good a priori reason for thinking that ratings will yield information of value before the personnel research worker feels justified in embarking upon an extended study of them as selection procedures.

## STEPS IN THE CONSTRUCTION AND ANALYSIS OF A NEW TEST

Whenever a program for personnel selection is to be based at least in part on new tests constructed specifically for that program, it is necessary to consider the operations involved in the construction, preliminary analysis, revision, and validation of a test. At this point, we shall outline the usual sequence of operations, commenting briefly upon certain of the steps in the chain. The particular statistical operations involved in pre-validation and validation analysis will be discussed in considerable detail in subsequent chapters. The usual sequence of steps will be approximately as follows:

1. Original conception of an idea for a test.
2. Development of specifications for construction of the test.
3. Construction of a preliminary test form.
4. Small-scale tryout of preliminary test form.
5. Statistical analysis of the preliminary form.
6. Preparation of a revised form.
7. Administration of the test for validation.
8. Determination of test's validity and correlations with other

measures.



### *Original conception of a test idea*

Little can be said about the original insight that suggests a test idea. The process of invention has never yielded very gracefully to psychological investigation. Without much doubt, the more a person knows about a particular job the more likely he is to have shrewd ideas about types of tests that will have validity for that job. By the same token, a wide knowledge of existing tests provides the background for fruitful new combinations of testing materials and procedures. The test inventor should, therefore, become intimately acquainted with the job or jobs for which the tests are being developed. He should also have a wide acquaintance with existing test forms and with the literature on the determination of distinct traits or dimensions of human performance. Beyond that, he must rely on his native wit and ingenuity.

### *Development of specifications for construction of the test*

The fact that the research worker has an idea for a test means immediately that he has in some vague form a mental picture of the finished test. This picture, vague at first, will become increasingly clear as work on the test proceeds. It will probably pay in most cases to make this mental picture as explicit as possible at each stage of the work by formulating a written statement of specifications for the test. This statement will clarify both for the test maker himself and for his colleagues just what he is trying to do in the test. The specifications might cover such points as the following:

1. The function or functions that the test is to measure.
2. Illustrations of each type of item to be included in the test.
3. The number of each type of item to be included.
4. The range of content to be covered in the test, where variety in content is a factor, and the allocation of items within that range.
5. The time limits for the test, or for each separately timed section of it.
6. The nature of the population for which the test is designed.
7. The desired level and range of difficulty of test items for the population.

8. The editorial and statistical procedures to be used in selecting and refining test items.

The specifications for a test must, of course, remain somewhat fluid while the test is in the preliminary stages of development. A description of a test covering the above points would represent the thinking about the test at that time. It would be subject to revision on the basis of new ideas concerning the test and particularly in the light of actual data from the trial of sample sets of instructions, sample items, or a rough and preliminary form of the test. The specifications must necessarily grow with the test.

A set of specifications has one value in crystallizing the plan for a test in the mind of its creator. It has a second value in making available to collaborators the explicit plan for the test, so that individuals may work together on it effectively and so that positive suggestions and criticisms may be stimulated and directed. A third and often quite important value of a set of specifications lies in its use as a guide to the construction of further equivalent forms of the test. If a clear statement is available of what a test contains in terms of type, content, and difficulty of items, and if the statement further specifies the procedures for item editing and selection, then the preparation of further forms of the test which are truly equivalent to the first form in functions measured, scope, and difficulty level should be greatly facilitated.

### *Construction and statistical analysis of the preliminary test form*

In the development of a new test, we can roughly identify three stages. These we shall speak of as (1) the exploratory stage, (2) the preliminary stage, and (3) the final stage. In the exploratory stage the test developer is concerned with working out the format for typical test items, with phrasing directions which will be uniformly understandable and will give each individual an opportunity to perform at his best level on the test itself, with determining roughly the rate at which the test items can be done, and generally with locating "bugs" in the test. At this stage, it pays to make frequent informal use of any small group of subjects who may be conveniently available as guinea pigs for try-out of the materials. Submission of the instructions

and perhaps some preliminary test items to available stenographers, clerks, students, or co-workers will provide some immediate information on the adequacy or, particularly, the inadequacy of the materials. Both the test performances of and the comments and criticisms from those who take the test will enable the author to identify and eliminate many difficulties before the test is reproduced in quantity and tried out on a wider audience.

When the difficulties revealed by the initial exploratory try-out have been overcome satisfactorily and informal work with individuals and small groups indicates that a feasible testing procedure has been developed, it will then be appropriate to develop a preliminary form of the test for administration to substantial groups of individuals having approximately the same characteristics as the population with whom the test is ultimately to be used. This administration is for the purpose of getting some preliminary statistical evidence about the test as a whole and about the characteristics of individual test items. The information about the test as a whole that is of particular interest at this stage is its reliability. The problem of obtaining estimates of the reliability or precision with which a test measures is discussed in detail in Chapter 4. Information is also often desired on individual differences in time taken to complete the test, so that appropriate time limits can be set for a final test form. For purposes of editing and revision, data are required on the difficulty level of the individual test items and the degree to which each item discriminates between those who are high and those who are low on total test score. The problems and uses of item analysis are expounded further in Chapter 8.

Where a preliminary test form is to serve as the basis for item selection and revision, more items must be included in the preliminary form than are to be included in the final form. The number of surplus items included depends somewhat on the type of test. Where items are of a rather simple and standard form, as in simple number problems or tests of word meaning, an excess of 25 or 50 per cent may provide enough to permit elimination of items that fail to discriminate or items not of appropriate difficulty. For types of items that are more complex and more difficult to write, such as "judgment" items, an excess of

100 per cent or even more may be needed to provide an adequate stock for the final test form.

### *Administration of the preliminary form*

When the preliminary form of the test is complete, it should be reproduced and administered to a sample of the same population with which it is ultimately to be used. If a sample of the exact population is not available, the sample studied should resemble it as closely as possible. The closer the resemblance, the more directly the sample statistics will be transferable to the population ultimately to be tested. To provide adequate item analysis data, the sample tested at this stage should preferably number four or five hundred. When the results from this preliminary testing are to be a basis for item analysis, the test should be given with quite ample time limits, so that most individuals will have a chance to try all items. If the test is primarily a speed test, it may not be desirable to plan an item analysis for it. Item analysis of a pure speed test is a relatively meaningless undertaking. However, if item analysis is planned for a test which will ultimately be administered as a speed test, the data for the item analysis should be obtained when the test is administered unsped.

### *Preparation of the revised form*

Analysis of results from the administration of the preliminary form of a test provides the basis for revising the test and preparing a form to be administered for purposes of validation. The results most extensively used in preparing the revised form are those from analysis of the single test items. These results provide a basis for selecting items from the available stock and for editing and revising defective items. Data on the reliability of the whole preliminary form of the test give some clue as to how long the final test needs to be to provide a reasonably precise measure of individual ability. Records of time taken to complete the preliminary form suggest the number of test items that should be allotted to a given amount of testing time. If the test is primarily a power test, this time allotment is set so that most of the individuals tested have an opportunity to attempt all or nearly all the test items. If the test is designed as a speed test,



time limits are set so that there are a few more items than the fastest worker can complete. Finally, the trial of the preliminary form under conditions of group testing may have revealed difficulties in instructions or in the mechanics of test administration of which the author had not previously been aware, and these must then be corrected in the revised form. The revised form constitutes, at least tentatively, the form ultimately to be used for selection purposes.

### *Administration of the test for purposes of validation*

When the test has been prepared in its revised and presumably final form, the next step is to determine its validity for the job or jobs being studied. The empirical determination of validity requires (1) the determination of a test score for each member of an experimental group, (2) the determination of a measure of success on the job for each member of this same group, and (3) the calculation of appropriate statistical indices of relationship between the test scores and the measures of job success. Criteria of job success will be discussed in Chapter 5. Indices of relationship between test and criterion will be considered in Chapter 6. Our present concern is with certain problems involved in getting the test scores for a group for whom criterion data will become available.

Practical problems in validation testing center around questions of whom to test, when to test them, and how many individuals to test. Shall we give our research test to all applicants when they first apply for a job? Shall we test only those who are accepted for a given job category? Shall we test experienced workers who have been on the job for a considerable period of time? Shall we test before or after a period of training for the job? The specific options and alternatives depend upon the unique features of the particular personnel situation. However, there are certain general considerations and alternatives which we can profitably discuss here.

There are two major and basically conflicting considerations in determining whom and when we shall test. On the one hand, we should like to have a minimum of waiting time and attrition in the tested population between testing and obtaining criterion data. From this point of view, it would be ideal to test experi-



enced personnel for whom records of success on the job are already available. On the other hand, we wish our validation testing to be carried out under conditions of motivation and experience as nearly as possible like those to be encountered in the final use of the test. From this point of view, our ideal would be to administer a research test to job applicants when they are applying and being considered for the job in question. The conflicting advantages of these two procedures bear some further elaboration.

In almost any personnel situation there must be a rather substantial time lapse between initial selection for the job and the maturing of usable criterion information about the individual. The employee must have time to learn the job and must perform it for enough time to yield a representative picture of his ability. Even in simple and routine jobs, this period of learning, adjustment, and building up of an adequate performance record may take several months. In jobs involving extensive special knowledges and skills the training period may take years. Thus, even under the pressures of the war emergency the minimum interval between the date of testing an AAF aviation cadet in order to estimate his aptitude for different types of air-crew duty and the date of completion of his individual flight training was often as much as a year. Many additional months of advanced tactical training were added to that before the man was assigned to combat duty, and many more months elapsed before a record of his combat performance became available. The period of apprentice training for such crafts as carpenter or plumber has been several years in length. And the training programs for most of the professions also extend for a number of years.

Of course, some criteria of success may be obtained for performance during the training period. In fact, it has been a rather general procedure to use such criteria, due to the great time lag introduced if one waits for actual records of success on the job. In wartime, new test development could not wait the two years or more that would have been required to obtain combat criterion data. In times of peace, relatively few personnel research projects have or are willing to take the time to follow up test results over a period of five or ten years to see how the candidate for

training in law, medicine, or accountancy makes out after he has had a chance to establish himself in his profession.

A second problem that arises if job applicants are tested at the time of their initial application is loss of cases between the time of testing and the time that criterion data mature. A portion of the applicants, perhaps a substantial portion, may be rejected on the basis of test scores or of some other consideration, and for them no criterion data ever become available. Others may decide not to accept the job or take the program of training. Others may drop out during the training period for one of a great variety of reasons, some related to success in the job and some entirely incidental to quality of performance. The problem of shrinkage in the criterion group becomes particularly acute when the tests are a basis for *classification* rather than *selection*. If the applicants being tested are candidates not for a single job but for one of several, then the number who will eventually be placed in any *one* job and for whom criterion data will become available for that particular job may be only a small fraction of the total tested. Thus, in the testing of air crew in the AAF, not more than five or ten of each hundred tested eventually entered training for the job of navigator or that of bombardier. To get a criterion group of 1000 navigation students, being satisfied with only a training criterion of success, it would have been necessary to test 10,000 to 20,000 men in classification centers.

These two factors of (1) time lag and (2) loss of cases between testing and the maturing of criterion data represent serious practical handicaps to the use of pre-tests of applicants as a basis for a personnel research program. From the point of efficiency in time and testing, the more attractive procedure would be to test individuals already on the job, for whom records of proficiency exist or can be procured without delay. Unfortunately, this procedure also has its practical drawbacks in many cases. It may be administratively difficult to assemble for testing a group of individuals who are busy in a job, both because of interference with the progress of the work and because of geographical factors. The main problems involved in testing individuals at these later stages, however, concern the interpretability of the test results. The interpretation of test results from on-the-job personnel may be questioned on two counts:

(1) Are the conditions of motivation and rapport for those tested on the job equivalent to conditions for job applicants? (2) How have test scores been affected by job training and job experience?

With regard to the first of these points, we may often question whether the same level and universality of effort can be obtained from those who are tested as part of a research project, and who see no direct relationship between the tests and getting a job, and those who are tested as part of the process of applying for a job or a program of training. Research tests may be taken grudgingly, and motivation may vary widely, depending on the individual's interpretation of the purpose of the testing, his feeling of adequacy in a test situation, and his general level of cooperativeness and affability. The level of motivation often will be lower and the variability in motivation usually will be higher if the test outcomes have no clear bearing upon the goals of the individuals tested.

In certain types of tests, we may expect training for and experience on the job to have an appreciable direct effect on the test scores. A test may be diagnostic of future success if applied at one level of experience and yet fail to be diagnostic at some other level. Thus in selecting airplane pilots a test of general information about planes and aviation was found to have appreciable validity as applied to applicants. This same test would certainly have given a very different score distribution and might have shown quite different validity if applied to pilots who had already completed some part of their training. Measures of skills, information, and interests related to the job will be most clearly affected by job training, but the possibility of job training transferring to test score is quite a real one in any test. This makes any validation based upon tests administered at some time during training or work on the job necessarily tentative.

The choice of a group and time for validation testing is at best a compromise. For the test results to be most clearly and directly applicable to the selection testing situation, research testing should be done with applicants at the time of their application and under the same conditions that will prevail in subsequent use of the test. On the other hand efficiency and prompt availability of results often suggest testing groups at more nearly the time that criterion data will be available. How-

ever, results obtained from testing during training or on the job necessarily provide only a tentative picture of the validity of a test as a selection device.

A decision as to the number of cases to be tested for validation purposes again presents a conflict between the theoretically desirable and the practically expedient. Ideally, the validation of a research test should be based on a large number of cases. In practice the number of available subjects, the amount of testing time, and the amount of time of research and clerical personnel are always limited. A number of research tests may be competitors for these limited facilities. Allocation must be made among the various competing research tests.

The need for large numbers of cases in validation arises out of the instability of the correlation coefficient from sample to sample. With a sample of 100 cases and a true population value for the correlation coefficient of 0.50, we may expect to get a correlation either as large as 0.61 or as small as 0.36 in 1 sample out of 10. The weights for combining tests to give the best joint prediction of a criterion are even more sensitive to fluctuations from sample to sample. In order to get an accurate index of the degree of relationship, therefore, and particularly to determine with confidence and precision the best combination of tests for predicting a criterion, test validities (and test intercorrelations) need to be based on substantial samples.

The appropriate size of validation groups depends on the one hand on the total flow of personnel into the job in question and on the other on the extensiveness of validation data already available for other tests. Obviously, we must be content with smaller validation groups for a job specialty that has an input of 200 men a year than for one in which the flow is 200 men a day. Even though we should have to spread any testing project over a much longer time span in the smaller group, we could never build up the substantial populations that we could in the larger. There comes a point where the flow is so small that personnel research is almost impractical, and the figure of 200 men a year certainly approaches that limit. Validity coefficients based on fewer than 100 cases are too unstable to serve as the basis for any serious comparative study of the validity of different tests for



personnel selection, and if six months is required to accumulate a population of that size a research program must progress slowly indeed.

The second consideration in deciding the size of populations for validation testing is the extensiveness of validation data already available as a backlog of information for certain tests. If we already have a pool of tests, some of which are routinely being used for personnel selection, and if we have validation statistics for those tests for substantial numbers of cases, we will require more evidence indicating the validity of a new test before adding it to or using it to replace some of the tests already in use. When a research program of personnel selection for a particular job is starting, it may be worth while to validate an assortment of tests on populations of two or three hundred, in order to obtain as quickly as possible a rough approximation of the validity of a number of tests. At a later stage, however, when a battery of tests has been used for selection for some time, and when repeated analyses of the validity of those tests have built up validation populations for many hundreds or even thousands, much more extensive data on any new test are required before it can receive serious consideration as a candidate for the battery.

In the AAF air-crew classification program, for example, the earliest validation testing was done with groups of a few hundred, and validity coefficients were obtained for groups of one or two hundred pilots. These data served a valuable purpose in providing some guide to the early composition of a testing battery. By the time this battery had been in use for two years, however, repeated analyses of the validity of the tests for the job of pilot provided data based on thousands and even tens of thousands of cases. The validities of those tests had been determined with great accuracy. At this stage of the program, it did not seem worth while to validate any new research test on groups of less than one or two thousand, because very convincing evidence would have been required for a research test to compete for testing time with one of the thoroughly tested tests in the battery.

In general, therefore, in the initial validation of a group of tests for a personnel research program, the time, facilities, and available flow of personnel should be divided among a number



of the most promising tests. As much testing should be done as possible, still permitting results to be available by the time at which a decision must be made. In a continuing test program, repeated validation studies should establish with greater and greater precision the validity of tests actually being used for selection or classification. As this precision becomes greater, larger groups will be required for each new research test, if it is to compete with the tests in the existing battery.

### PREPARATION OF OBJECTIVE TEST ITEMS <sup>10</sup>

The backbone of most mass testing programs is the printed test made up of a substantial number of separate brief items. The standard forms of the short-answer test question are already quite familiar to anyone who has gone to school or otherwise had occasion to take tests in recent years. A single item of knowledge, formulated in various ways, will serve to illustrate these. Some of the most familiar types are the following.

#### *True-false*

True   False   The first president of the United States was George Washington.

#### *Multiple choice*

(   ) Who was the first president of the United States?

- A. Benjamin Franklin
- B. George Washington
- C. John Adams
- D. Thomas Jefferson
- E. Abraham Lincoln

<sup>10</sup> This section is a brief and necessarily superficial presentation of the whole problem of test item writing. The reader who is interested in further help on this problem is referred to the standard texts on educational and psychological measurement, and especially to the following:

D. C. Adkins et al., *Construction and Analysis of Achievement Tests*, U. S. Government Printing Office, Washington, D. C., 1947.

H. E. Hawkes, E. F. Lindquist, and C. R. Mann, *The Construction and Use of Achievement Examinations*, Houghton Mifflin Co., Boston, 1936.

Technical Staff, Board of Examinations, University of Chicago, *Manual of Examination Methods*, University of Chicago Bookstore, Chicago, 1937.

R. W. Tyler, *Constructing Achievement Tests*, Ohio Bureau of Educational Research, Ohio State University, Columbus, Ohio, 1934.

*Matching*

- |       |   |                       |
|-------|---|-----------------------|
| _____ | 1. First president of the United States | A. Alexander Hamilton |
| _____ | 2. Chief Justice of the Supreme Court   | B. John Marshall      |
| _____ | 3. Secretary of the Treasury            | C. Patrick Henry      |
|       |   | D. George Washington  |
|       |   | E. Andrew Jackson     |

*Completion*

The first president of the United States was named \_\_\_\_\_.

*Unrestricted response.*

Who was the first president of the United States?

The advantages of a test made up of items in these forms lie in the breadth of sampling of behavior which the test provides, the objectivity which is possible in evaluating individual performance, and the speed and convenience with which the test may be scored. Breadth of sampling results from the number of separate items which it is possible to include in a limited amount of testing time. Objectivity and ease of scoring are automatically favored by the brevity of the required answer. However, not all short-answer tests are equally objective or easy to score.

Objectivity stems from uniqueness in the correct response. Only when the correct response can be specified as some one number, name, or the like does evaluation of the response become strictly objective. This situation prevails most clearly when the subject being tested must select the best answer from a given set of response alternatives. When the testee produces the response rather than selects it from an established set of alternatives, a very large number of different responses may be obtained. These may often be graded by small steps from the clearly correct response planned by the test maker through approximations to it and on to responses that are clearly incorrect. Judgment is then required to determine how closely the response given corresponds to the ideal and whether credit should be given for it. The more difficult the items become, the higher is the level of ability required in the scorer in order to make the necessary discriminations.

The scoring becomes completely objective in such test forms as true-false, multiple choice, or matching. In these, once a

scoring key has been established on the basis of a consensus of experts, the one correct response for each item is entirely specified. Such a rigid scoring key not only permits the test to be scored by clerical personnel with no special knowledge or judgment in the area concerned but also makes the test suitable for mechanical scoring, using such a device as the IBM Test Scoring Machine. Where testing is on a scale sufficiently large to justify the expense, machine scoring usually speeds the scoring of tests very markedly and improves scoring accuracy.

The familiar forms of completely objective items are open to criticism because they call for a limited type of recognition solution of the task presented by the items. This criticism is especially relevant in achievement and proficiency tests, where validity is judged by the closeness with which the test task duplicates the actual functions performed on the job. Usually the real task on a job is to *produce* the correct response, not to *select* it from a few alternatives already provided. One outlet for ingenuity in test construction lies in the designing of item types that require the testee to produce the response but still retain the objectivity of a unique correct answer and permit the convenience of machine scoring. Two examples of this sort may be cited.

The conventional vocabulary test item is of the following type:

Which word means most nearly the same as *tepid*?

- (1) Tent shaped
- (2) Tilted
- (3) Warm
- (4) Sweet
- (5) Spoiled

In the American Council on Education Psychological Examination, the cues presented to the subject have been somewhat reduced and the items have been constructed differently, thus forcing the subject to depend to a greater extent upon active recall. The following is an example:

A four-letter word meaning the same as *tepid* begins with the letter:

- (1) B (2) N (3) S (4) T (5) W

In the AAF a test requiring the reading of numerical scales was arranged so that a standard code was used for recording the *last digit* of the answer. An answer sheet with space for ten different responses was used. Any answer ending in 0 was to be marked in the *A* position on the answer sheet, an answer ending in 1 in the *B* position, and so forth. In this test, any error was almost sure to occur in the final digit, so that this procedure gave the subject complete freedom of response but still caught practically any error. Similar devices can be applied in other tests to limit the cues provided the testee and yet permit complete objectivity in scoring of his responses.

Although writing good test items is to a considerable extent an art, consideration of the factors to be looked for in an item may guide the novice through some of the more common pitfalls. This discussion will center on the multiple-choice type of item, which is by far the most common item type in well-constructed objective examinations. The multiple-choice item consists of two parts, the stem and the response options. The stem of the item states the problem. In our earlier example it is "Who was the first president of the United States?" The response options are the four or five choices from which the examinee must select the best one. In the example these are the five names.

As its first consideration a test item should cover content that is significant for the purposes of the test. This is particularly true of achievement or proficiency tests. Thus, if a test of weather knowledge is being prepared for airline pilots, each test item should be scrutinized to make sure that it deals with an element of weather knowledge that is really of concern to the pilot. The judgment of specialists in the field is of the greatest importance at this point. In aptitude tests, the question of appropriateness of content is less readily determined. However, the function underlying the test should be kept in mind as each item is developed.

Since most tests are not intended to measure ability to understand difficult and obscure writing, the linguistic difficulty of test items should ordinarily be kept low. Except in tests of reading ability and verbal comprehension, the language of test items should be kept so simple that very few examinees fail the item because they do not know the words used in the question or

because they cannot unravel the complexities of the sentence structure. This problem is a very real one for tests of arithmetical problem solving, mechanical ability, social intelligence, and a variety of other functions. The liberal use of pictures or diagrams often provides an appropriate technique for minimizing language difficulty.

There are two goals in preparing response options for multiple-choice items. In the first place, the options should be brief, with as much of the item as possible incorporated into the stem. An item would not be written:

The first president of the United States

- A. was named Benjamin Franklin.
- B. was named George Washington.
- etc.

but rather

The first president of the United States was named

- A. Benjamin Franklin.
- B. George Washington.
- etc.

The second form reduces the amount of reading material included in a single item, and also simplifies somewhat the expression of the item.

The second goal in selecting response options is that the misleads should be plausible and attractive. In order to justify its inclusion, each option must be chosen an appreciable number of times by individuals of limited ability in the function being measured. In the selection of misleads, the purpose is to pick options which are definitely not right, but which will attract the individual who does not know with assurance the answer to the problem presented by the item. Devising these response options requires a great deal of skill. It is an art to decide: How would the inept person be likely to answer this question? Where time and facilities permit, this art may be supplemented by preliminary experimentation with the item in completion or free-answer form. The most common errors resulting from free response to the item provide the logical options for use when the test is cast into multiple-choice form. However, some studies suggest that



the insights of the skilled writer of items are at least as valid as free-response data as a basis for multiple-choice options.

One vital step in the preparation of good test items is the editorial review. The review serves the two purposes of locating ambiguities in statement and of checking the authenticity of the key. It is an axiom of writing that the author of a statement is not best qualified to judge the clarity of the statement to others. The author knows the idea that he is trying to convey; therefore he cannot appreciate the difficulties that others will have with his idea. He can achieve some perspective on his work by reviewing it after a period of time. The best check of its clarity to another, however, is having it read *by* another. The independent review by an editorial reader provides an effective source of criticisms and suggestions for refining the test item. At the same time, an independent judgment of the keying of test items is always desirable. If the material of the test is technical or deals with a subject in which the item writer does not have specialized competence, review by an expert becomes imperative. It is only through such a review that one can determine that there actually is one and only one correct answer to the item. If a test is to retain the respect of specialists in the field with which it deals, the items must be correctly and unambiguously keyed.

One particular fault the item writer must avoid in preparing test items is providing irrelevant cues or "specific determiners" for the correct response. In their most obvious form, these cues arise out of the grammatical structure of the question. Thus an item might read:

A man who designs houses is called an

- A. contractor.
- B. architect.
- C. carpenter.
- D. engineer.
- E. plumber.

In this example, the form "an" in the stem automatically eliminates three of the response options. The cue in this case is very obvious. In other cases, the extraneous cues may be more subtle and indirect. In true-false questions, for example, the words "all" or "none" in the statement are probable cues that the state-

ment is false, whereas "many" or "few" tend to characterize statements that are true. Also, in true-false questions, long statements with many qualifying clauses are usually keyed as true, whereas shorter statements are more likely to be false. If an item can be solved by cues of this sort, the purpose of the test is usually defeated.

When the items for a test have been prepared, they must be assembled into a test form and reproduced in quantity for administration. The preliminary form of the test will usually be needed only in fairly limited quantities, and it is often satisfactory to use stencil or gelatin duplicators. A much better quality of reproduction can be obtained by photo-offset or by actual printing. These processes are, of course, much better suited to the preparation of large editions of a test.

In assembling a test there are a number of factors to be considered. Attention may first be turned to the instructions. It is very important that these be clear and adequately detailed. When the test is of a familiar form and the procedures are simple, a brief paragraph of instructions may suffice. Whenever the testing procedure promises to be at all novel for the group to be tested, however, instructions should be full and detailed. They should include illustrative examples and practice examples, and it is often desirable to provide time for a brief practice test. Instructions should be pre-tested with individuals and small groups comparable to the less able individuals in the groups eventually to be tested. They should be revised on the basis of this preliminary exploration until they prove satisfactory.

The sequence of test items is a second consideration. In a speed test, items are presumably all easy and all more or less equivalent, so that sequence is a minor matter. In a test with a definite power element it is good practice to arrange items in an approximate order of difficulty, starting with the easiest items. Items referring to a single passage of text, diagram, or table should, of course, be grouped together. Also, it is usually desirable to group all items which are to be answered in the same way or which depend on a single set of instructions.

The format and typography of a test should be based on considerations of legibility, convenience in taking the test, convenience in scoring, and attractiveness. In the interests of

legibility and attractiveness the test form should not be too crowded. Type should be of good size and diagrams and illustrations should not be cramped. Short lines of text are more legible, and a test often appears better if it is printed two columns to the page. A test is more convenient to take if the need to turn pages is reduced to the minimum. A single item should not run over from one page to the next, and if possible all the items referring to a particular passage of text, diagram, or table should be on a single page or on facing pages. It will then not be necessary to turn back and forth from question to diagram. This makes the test more convenient for the user and also lessens the possibility of error on his part because of losing his place. Scoring is easiest if a separate answer sheet is used upon which answers to all items are entered. The separate answer sheet is usually a necessity for machine scoring, but it also lends itself very well to hand scoring with a stencil. The separate answer sheet makes it possible to use a test booklet a number of times. This provides a substantial economy in a continuing testing program. If answers are entered upon the test blank itself, spaces for answers should be placed in a column, usually on the right-hand side of the page. A scoring stencil may then be laid alongside the answers, and the responses may readily be compared with the key. A neat and attractive cover page helps any test blank to make a good impression.

## *The Estimation of Reliability*

The two qualities of a single personnel selection test in which the research worker is particularly interested are its reliability and its validity. A measurement procedure is reliable to the extent that repeated measurement gives consistent results for the individual—consistent in that his score remains substantially the same when the measurement is repeated, or in that his standing in the group shows little change. A measurement procedure is valid in so far as it correlates with some measure of success in the job for which it is being used as a predictor.<sup>1</sup> The qualities of reliability and validity are important in each single test. As soon as two or more tests are used as joint predictors of a criterion of job success, the research worker also becomes concerned with the correlation between the separate prediction tests. The manner in which correlation between tests enters into the prediction problem will be considered in Chapter 7, in which the development of a battery of tests for prediction is discussed. Test validity will be the subject of Chapters 5 and 6. The present chapter deals with the concept of reliability and the uses and limitations of data on reliability for evaluating personnel measurement procedures.

Reliability or consistency in a measurement procedure is a matter of degree and not an all-or-none matter. Whenever we measure anything, whether in the physical, the biological, or the social sciences, that measurement contains a certain amount of chance error. The amount of chance error may be large or small, but it always is present to some extent. If the chance errors are small in size, relative to the variation from person to person, the reliability or consistency of the measure is high. If

<sup>1</sup> This definition of validity is somewhat limited. For a fuller discussion of the topic see Chapter 5.

the chance errors become large in proportion to the variation from person to person, the reliability of the measure is low.

In evaluating the consistency of a set of measurements, there are two somewhat different aspects to be considered. These may be spoken of as *absolute* consistency and *relative* consistency. The degree of absolute consistency is seen in the actual amount of variation which results when a particular measuring instrument is applied more than once to the same individual. The standard deviation of such a distribution of repeated measurements, called the *standard error of measurement*, is the statistic used to express the absolute consistency of a measurement procedure. For the evaluation of relative consistency, some statement is required of the degree to which individuals keep the same relative standing in the group when two equivalent forms of a test are applied to all the members of the group. The correlation between the two sets of scores, called the *coefficient of reliability*, provides an index of consistency relative to the group as a whole. Although absolute consistency is more meaningful if a descriptive statement of the accuracy of measurement is desired, practically all analytical study and critical comparison of tests are based on the measures of relative consistency.

### LOGICAL CONSIDERATIONS IN EVALUATING RELIABILITY

The evaluation of the reliability of a measuring instrument involves two types of operations, one experimental and the other statistical. On the one hand, it is necessary to apply the instrument to a defined group of cases following a specified experimental plan and maintaining specified experimental conditions. On the other, the scores resulting from such administration must be analyzed by appropriate procedures to yield a statistic which will represent the reliability characteristics of the test. These two aspects are somewhat independent, in that essentially the same statistical procedures may be applied to data gathered in quite a variety of ways.

The experimental procedures underlying an estimate of reliability are very closely bound up with the logical aspects of the problem, so that one needs first to analyze the purposes to be



served by a measure of reliability. The experimental operations must be planned and evaluated with these purposes in view. The next sections of this chapter are organized around the analysis of the logical and experimental aspects of reliability. Consideration of statistical procedures follows discussion of the various experimental procedures, a given statistical analysis being discussed in connection with the experimental procedure with which it has the closest connection.

### *Reliability and analysis of variance*

Whenever a measuring device is applied to a group of individuals and a score is obtained for each individual in the group, the resulting distribution of scores will spread out over an appreciable range of score values. The variation in any set of scores arises from a number of different contributing factors. Some of these factors refer to true differences among the individuals in the group with respect to the quality being measured. Some represent sources of inaccuracy in the measurement of the separate individuals. The evaluation of the reliability of any measure reduces to a determination of how much of the variation in the set of scores is due to true differences among the individuals in the group and how much to inaccuracies in measurement of the particular individuals.

A number of different statistics have been developed as summary values to describe the variability in a set of scores. For the purposes of the present discussion, the most useful will be the variance  $\sigma^2$ .<sup>2</sup> This is defined as the average of the squared deviations from the average of the group.

$$\sigma_x^2 = \frac{\Sigma(X - M)^2}{N} = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2 = \frac{\Sigma x^2}{N}$$

The particular advantage of the variance for the present discussion is that it can be broken down into separate parts which combine additively to give the total. Thus, if the variance of weight, in pounds, of pupils in a class was 150, this might break up into a variance of 125 permanently associated with the individuals and a variance of 25 associated with the accidents of

<sup>2</sup> In representing the variance of a distribution, the symbol  $\sigma^2$  will represent the theoretical population value, and  $s^2$  will designate the value obtained from a specific limited sample.

that particular set of measurements. These parts added together make up the total variance of 150 for the set of scores. Whenever a number of independent factors combine to produce a score, it is possible to make an analysis of variance into fractions which are associated with particular factors. Conversely, these fractions add up to the total variance.<sup>3</sup> That is,

$$\sigma^2 = \sigma_a^2 + \sigma_b^2 + \dots + \sigma_k^2$$

where  $\sigma^2$  is the total variance of the distribution of scores and  $\sigma_a^2, \sigma_b^2, \dots, \sigma_k^2$  are the fractions of the variance associated with factors  $a, b, \dots, k$ , respectively. Thus, the variance in weight of pupils in a classroom might be broken down into variance associated with age, with sex, with family, and with each other definable characteristic of the individuals in the group. There would also be variance associated only with the one particular set of measurements, that is, variance that would not be reproduced another time. This may be designated *error variance*. The existence of this error variance corresponds to the fact of unreliability, and its amount relative to the total of all variance is a measure of the degree of unreliability.

Let us designate the variance of true scores of a group on a trait by  $\sigma_\infty^2$  and the variance of errors of measurement by  $\sigma_e^2$ . If the error of measurement is independent of the true score, we have

$$\sigma^2 = \sigma_\infty^2 + \sigma_e^2$$

That is, the variance of the obtained scores equals the variance of the true scores plus the variance arising from errors of measurement. It is also possible to relate these fractions of variance to the reliability coefficient discussed earlier in the chapter. We have

$$r_{11} = \frac{\sigma_\infty^2}{\sigma^2} \quad (1)$$

and

$$r_{11} = 1 - \frac{\sigma_e^2}{\sigma^2} \quad (2)$$

<sup>3</sup> When factors are not independent, it becomes necessary to analyze covariances as well as variances.

That is, the numerical value of the reliability coefficient of a test corresponds exactly to the proportion of the variance in test scores that is due to true differences between individuals in the quality being evaluated by the test. A test is unreliable in proportion as it has error variance.

It becomes clear that the basic problem in determining the reliability of any measurement procedure becomes that of *defining* what shall be considered true variance between individuals and what shall be thought of as error variance. When this definition has been reached, the next step is to devise a series of experimental and statistical operations which will provide the best estimate of the defined fractions of variance.

### *Sources of variance in test scores*

As noted above, variance in a set of scores from any test or measuring device arises from a great variety of specific sources. However, these may profitably be grouped, for purposes of discussion, into a few major categories. A classification of sources of variance is presented in Table I. The categories given here probably do not exhaust the possible range of categories. Certainly, many more subcategories could be listed under most of the major headings, and the list presented should be thought of as illustrative rather than exhaustive. A consideration of each of the categories will provide the basis for a decision as to which fractions of variance should be thought of as true, systematic variance in the quality or qualities being measured and which should be thought of as error variance.

Variance within a set of scores arises first of all because different individuals possess different amounts of certain general and persistent traits (category I of Table I). Thus some type of ability to reason deductively might be a general quality which entered into performance on a number of tests of intellectual functioning. Verbal comprehension is likely to enter into a wide range of tests requiring reading. Almost any test performance depends in part on general abilities which are also involved in a number of other types of test performance. Since this variance is due to a lasting characteristic of each individual, it is clearly systematic variance and should be treated as such in any sequence of operations set up to provide an estimate of reliability.

TABLE I. POSSIBLE SOURCES OF VARIANCE IN PERFORMANCE ON A PARTICULAR TEST

- I. *Lasting and general characteristics of the individual.*
  - A. Level of ability on one or more general traits, which operate in a number of tests.
  - B. General skills and techniques of taking tests.
  - C. General ability to comprehend instructions.
- II. *Lasting but specific characteristics of the individual.*
  - A. Specific to the test as a whole (and to parallel forms of it).
    1. Individual level of ability on traits required in this test but not in others.
    2. Knowledges and skills specific to particular form of test items.
  - B. Specific to particular test items.
    1. The "chance" element determining whether the individual does or does not know a particular fact. (Sampling variance in a finite number of items.)
- III. *Temporary but general characteristics of the individual.* (Factors affecting performance on many or all tests at a particular time.)
  - A. Health.
  - B. Fatigue.
  - C. Motivation.
  - D. Emotional strain.
  - E. General test-wiseness (partly lasting).
  - F. Understanding of mechanics of testing.
  - G. External conditions of heat, light, ventilation, etc.
- IV. *Temporary and specific characteristics of the individual.*
  - A. Specific to a test as a whole.
    1. Comprehension of the specific test task (in so far as this is distinct from I B).
    2. Specific tricks or techniques of dealing with the particular test materials (in so far as this is distinct from II A 2).
    3. Level of practice on the specific skills involved (especially in psychomotor tests).
    4. Momentary "set" for a particular test.
  - B. Specific to particular test items.
    1. Fluctuations and idiosyncrasies of human memory.
    2. Unpredictable fluctuations in attention or accuracy, superimposed upon the general level of performance characteristic of the individual.
- V. *Variance not otherwise accounted for (chance).*
  - A. "Luck" in the selection of answers by "guessing."

Two rather special types of persisting general factors deserve particular mention. These are general ability to comprehend instructions and what we may speak of as "test-wiseness." These factors are mentioned because they are likely to enter into any

test score, whether we want them to or not. Performance on many types of tests may be in some measure a function of the individual's ability to understand what he is supposed to do on the test, particularly if the test situation is novel or the instructions are complex. At the same time, the test score is probably in some measure a function of the extent to which the individual is at home with tests and acquainted with tricks of taking them. The presence of variance in score due to variation in comprehension of instructions and in "test-wiseness" is frequently undesirable for the purposes of the test in question. However, these factors must be recognized. They present a challenge to the author of the test, who will try to minimize them, except where their presence is specifically desired, by providing the clearest possible instructions and a minimum of secondary cues. These factors present a practical rather than a theoretical problem; statistically they represent a general, lasting quality of the individual and must be treated as such.

In addition to the variance common to a range of tests, each test has some variance arising from the persistent characteristics of the individuals being studied yet specific to the particular area being tested (category II). That is, some variance will be present in spelling tests, for example, but not in tests of any other performance. There are, of course, degrees of specificity of knowledge or skill, so that further narrowing down may take place even within a given field. In addition to variance characterizing the field of performance, such as spelling or numerical computation, there may be variance associated with the specific form and manner of testing. In a spelling test this might depend on whether the test consists of oral presentation as in a spelling bee, writing words from dictation, or recognition of errors in words presented in a printed test. Finally, in any test there is usually variance associated with the particular sample of test items. Given two tests made up of samples independently chosen in the same way from the same universe of items, an individual will fail to receive identical scores on the two tests because of variation in the particular items for which he happens to have the necessary skill or information.

At this point we begin to encounter difficulty in deciding which fractions shall be included in the systematic variance and which



thought of as error variance. Variance specific to the area covered by the test (category II A 1) is certainly systematic variance, and any operations for determining reliability should be so planned that this type of variance is treated as systematic. The problem arises in connection with variance associated with the particular test format and with the particular sampling of test items. The question is to determine the most useful definition of reliability. Shall we define it in terms of a subject or a function only? In terms of a function and a specific form of test? Or in terms of a function, a form of test, and a particular set of test items?

Defining reliability in terms of function alone leads to experimental procedures that appear to come closer to evaluating test validity than test reliability. That is, if format and item type are considered sources of error variance, we are led to correlate tests with different types of items and manner of presentation. We are then beginning to inquire whether the test is consistent with other measures rather than whether that particular test measures consistently. Defining reliability in terms of function, form of test and specific set of test items is so narrow that it has little practical meaning in most cases. We are rarely interested in performance on a limited set of test items for their own sake. We are almost always interested in test performance as an indication of ability to perform on the whole universe of items of which the test represents a limited sample. The variance due to the sample of items is, therefore, in an entirely true sense part of the error of measurement. In conclusion, then, in the most meaningful definition of reliability, variation arising from the specific abilities required in the area being studied (category II A 1) and from knowledges and skills specific to the particular form of the test (category II A 2) is treated as systematic variance, but variation in performance arising from the particular sampling of test items (category II B) is treated as error variance.

The above discussion emphasizes one very important point. There is no single, universal, and unequivocally correct reliability coefficient for a test. Determination of reliability is as much a logical as a statistical problem. The appropriate allocation of variance from different sources calls for practical judgment of the ultimate use to be made of the resulting statistical value.

This point will become increasingly apparent as the discussion continues.

A third group of factors producing variation in test scores are certain general but temporary characteristics of the individual or of the testing situation. These include state of health, amount of sleep the previous night, presence or absence of worries or other distracting influences, weather, ventilation, illumination, and a host of other internal and external factors which may affect the individual's mood and the efficiency of his work. Different test performances are susceptible to these factors in varying degrees, but all are probably influenced by them in some measure. The factors vary both in their permanence and in their generality. Some may change from day to day, some from hour to hour. There may even be very short time fluctuations in efficiency which represent changes from minute to minute. In general, however, the present category includes factors which characterize a particular testing session but not another session.

Here, again, a problem arises as to how to allocate variance of this type. Once again, the difficulty is to decide what type of consistency should be measured. Is it important to determine how consistent a measure we have of the individual as he exists at a particular moment? Or is it important to determine how consistent his performance is from day to day and week to week? There may be purposes for which the former is the significant information. Thus, in making an experimental study of day-to-day fluctuations in mood, we might be interested in knowing how reliable our assessment of mood on a particular day was. In general, however, in personnel psychology we are interested in a test score as representative of an individual's level of performance not on a specific day but over some period of time. Factors that cause his score to vary from one day to the next are usually sources of error as far as the purposes of our measurement are concerned.

Our discussion so far has provided no indication of whether this or any other *possible* source of variance does in fact yield practically significant amounts of variance. That is, we have not shown whether or not it makes any practical difference what we do with variance in the above category. That cannot be a matter of general theoretical discussion but must be a matter of specific

empirical evidence. The answer will probably vary widely in different areas of measurement. For example, in a simple power test of vocabulary, very little of the variance would be accounted for by temporary characteristics of the individual, but in a test of body metabolism those factors would probably be the source of a substantial amount of variation.

A further group of factors contributing to variation in test performance consists of certain relatively temporary and specific factors. In this category are included influences which tend to be more limited both in time and in scope than those discussed in the immediately preceding paragraphs. Certain of these factors characterize performance on a test as a whole. If the test is novel and the instructions difficult, individuals may vary in the extent to which they "catch on" to the nature of the task. In part this probably represents general ability to understand instructions (category I B), but in part it may represent temporary or "chance" variations superimposed on that general ability. Again, a test may have certain specific tricks or techniques of which the individual does or does not "get the hang." Furthermore, performance on many tests, particularly measures of complex coordination or skill, is susceptible to considerable improvement through practice. A temporary feature of some importance may be the individual's practice level at the moment of testing. Finally, there are certain factors which, for the lack of any better term, we may group together under the heading of "mental set" at the time of taking the test. Was the subject emphasizing speed or accuracy if it was a speeded test? To what cues was he particularly alert if it was a perceptual task? What was his momentary mood if it was an attitude or interest questionnaire?

The factors grouped in this category (IV A) are the ones whose presence and significance in any given case are probably most open to question. In many types of simple and standard tests they can very likely be ignored. However, in novel types of tests, highly speeded tests, measures into which introspective interpretation enters heavily, and perhaps other types, the possibility of encountering variance from such sources must be considered. This variance should probably be treated as error variance in most personnel research.

In addition to factors specific to a particular test and date of testing, there may be even more specific and temporary factors. These are factors which are specific to an item or a few items and a minute or a few minutes of time. They include short-time fluctuations of memory or attention, momentary blockings of performance, cyclic variations in effort, and a variety of other fluctuations superimposed upon the general level of performance. These factors (category IV B), as far as they affect score, introduce variable and unpredictable error into the score, and in the treatment of results this variance should be allocated to error.

Finally, we must introduce the concept of "chance" to take care of variance not otherwise accounted for. We can never find antecedent factors to account for all the variance in a set of test scores. Some variance arises from guessing at answers, some from other obscure variable influences which we cannot define or specify. The variance of this type (category V) is error variance in its purest form and the operations which define reliability must allocate it to that category.

### PROCEDURES FOR ESTIMATING RELIABILITY

The evaluation of the reliability of a measuring instrument requires a determination of the absolute or the relative consistency of repeated measurements of the same object or group of objects. In the physical sciences many repetitions of a measurement of a single object or phenomenon may provide a reasonable method for estimating the precision of the measurements. In dealing with human behavior, however, the individual is likely to be changed as a result of the operation of measurement, and it will usually be necessary to limit sharply the number of times a single individual is measured. In practice, therefore, all procedures of reliability determination generally useful in personnel research are based upon getting a small number of measurements, typically only two, for each individual in a representative group. Stability of results is achieved by increasing the number of individuals measured rather than the number of measurements of each. These measurements provide sets of scores, again usually two for each individual, for analysis. The usual analysis

consists of computation of the coefficient of correlation between the two sets of scores.

We have defined the reliability coefficient as the correlation between two sets of *equivalent* measures of a characteristic for a group of individuals. We must next consider what actual testing operations correspond satisfactorily to the logical requirements for equivalent tests. A number of different testing and statistical procedures have been proposed to provide the necessary coefficient of correlation between equivalent measures. We shall consider the major sets of experimental and statistical procedures in turn, describing and evaluating each in terms of its treatment of different categories of variance. The major procedures are:

1. Administering two equivalent tests and correlating the resulting scores.
2. Administering the same test form or testing procedure twice and correlating the resulting scores.
3. Subdividing a single test into two groups of items, each scored separately, and correlating the resulting two scores.
4. Analyzing the variance among the individual items of a test, and determining the error variance therefrom.

A section will be devoted to each of the above procedures, describing various subprocedures and discussing the problems of each.

### *1. Reliability defined by equivalent test forms*

Since the formal definition of reliability has been phrased in terms of the correlation between two equivalent sets of measures, it follows that the procedure for reliability determination that makes use of two equivalent tests will measure up to our logical requirements. We must, however, establish some satisfactory set of operations for preparing truly equivalent tests. This is a problem in the logic and practice of test construction. In preparing equivalent test forms one danger is that the two tests will vary so much in content and format that each will have some specific systematic variance (category II A) distinct from the other, in which case the correlation between the two will underestimate the reliability. Conversely there is the danger that the two forms may overlap to such an extent in specific details of



content that variance due to specific sampling of content (category II B) may be common to the two tests. In that case, this variance will be treated as systematic rather than chance variance, and the obtained correlation will overestimate the reliability.

The best guarantee of equivalence for two test forms seems to be a complete and detailed set of specifications for the test, as described in Chapter 3, prepared in advance of the assembly of any final test form. The specifications should indicate item types, difficulty level of items, procedures and standards for item selection and refinement, and distribution of items with regard to the content to be covered, specified in as much detail as feasible. If each test form is then built to conform to the outline, while at the same time care is taken to avoid identity or detailed overlapping of content, the two resulting test forms should be truly equivalent. That is, each test must be built to the same specifications, but within the limits set by the complete specifications each test should present a random sampling of items.

Two tests thus constructed will treat as systematic the variance in categories I and II A of Table I. They will treat as error variance that in categories II B, IV B, and V. The allocation of variance in categories III and IV A depends on the time interval between the administration of the two forms. If they are given in immediate sequence, this last variance is treated as systematic variance; if some time intervenes between the testings, this variance is allocated to error. For most uses of the resulting statistic, it is probably more meaningful to let some time elapse between the two testings, thus treating temporary day-to-day fluctuations as errors of measurement. The question of length of the interval arises. For most purposes, the answer depends on the fact that it is day-to-day fluctuations which we wish to allocate to error. An interval of a few days or weeks is sufficient. With longer intervals the problem of genuine growth and change in the individual is encountered, and the coefficient may be lowered because of these changes. Of course, for some purposes we may be interested in consistency of performance over an extended period of time, but consistency of this type probably goes beyond the concept of reliability.

In most of the usual types of tests of ability or achievement, preparing and administering equivalent forms should not present

undue technical difficulty. There are some situations, however, in which equivalence is very difficult to achieve. This is true when either (1) the test task is essentially unique or (2) a single exposure to the test changes the individual to such an extent that he is really a different individual at the second exposure. A unique test task occasionally occurs in connection with unusual problem-type tasks. Changes in the individual are a more common source of difficulty. In any task that is sufficiently novel so that the experience of being tested adds a significant increment to the individual's practice with the task, he becomes a somewhat different individual at the time of a subsequent test. In novel tasks, or in tasks which present essentially a learning situation, the changes may be quite marked. The problem of defining reliability for such a changing function is a very difficult one, and no completely satisfactory solution seems available.

## *2. Reliability defined by repetition of the identical test form*

In some cases, obtaining two equivalent measures reduces to repetition of the identical measuring instrument, the only difference being the time at which it is administered and perhaps the person by whom it is administered. Thus, two equivalent measures of weight can be obtained by weighing the members of the group being studied on the same scales at two times, ten days apart. The same situation holds for almost any physiological or anatomical measurement. In these cases, we do not encounter the problem of sampling items from a larger universe of behavior, and therefore distinct equivalent test forms are neither meaningful nor possible. Equivalence in this case means identity of measuring instrument and procedure.

There are also certain behavioral measures in which the situation of sampling from a large universe of items does not arise. This is the case in simple repetitive tasks of motor speed and skill or of perceptual judgment. Thus, in a test of simple reaction time, in which a measurement of the individual is obtained by timing repeated simple reactions to some stimulus, the test task is so defined that no varied sampling from a more extensive universe of behavior is involved. Here, again, repetition of the

same test provides the meaningful definition of "equivalent measures." The same is true of the type of perceptual judgment that is involved in a simple psychophysical experiment, as with judgments of brightness, length, weight, and so forth. In all such tests, repetition provides an acceptable procedure for estimating reliability.

In most measures of intellect, temperament, or achievement, however, repetition of the same test form and correlation of the two resulting sets of scores is less defensible as an operation for determining reliability. In these cases, a particular test consists of a limited sample from a much larger universe of possible items. The test score has practical significance in so far as it is representative of the individual's ability to respond to all the tasks in the universe from which it undertakes to sample. Reliability is a matter of the adequacy of the sampling of items as well as the consistency of behavior by each individual. In other words, in this case sampling of items (category II B) is an appreciable source of variance. Practical usefulness for the result dictates that this variance should be treated as error variance in determining the reliability of the test. Repeating the same test form holds the sampling of items constant so that this factor is treated as systematic rather than error variance. Reliability coefficients calculated from a repetition of the same test may be expected to be higher than those based on parallel, equivalent forms by an amount that is equal to this variance associated with sampling of items.

A second possible difficulty with repetition of the same test is actual memory of particular items and of the previous response to them. If this memory is effective in leading the individual to repeat the same response he made the time before, the results on two test administrations tend to be abnormally alike. The same answers may be repeated not because the individual is consistent in his behavior and arrives at the same conclusion in the same way, but because he happens to have a memory of his previous response. In effect, some of the variance associated with momentary memories and chance choices (categories IV B and V) becomes common to the two testings and is treated as systematic variance. Memory of previous responses is likely to be a factor

in proportion as (1) the test is short, (2) the test items are distinctive and memorable, and (3) the interval between testings is short.

In summary, for those types of test in which sampling of items and memory of previous responses are not an issue, a second application of the same test at a later date and correlation of the two sets of scores provides an adequate set of operations for reliability determination. In the many other tests, however, in which the factors of sampling and memory are significant sources of variance, repetition of the same test form will yield an estimate of reliability which tends to be systematically too high. In these latter cases the procedure is to be avoided.

### *3. Reliability defined by subdivision of a single total test*

The preparation and administration of two equivalent test forms, though logically entirely satisfactory as a procedure for determining reliability, present certain practical difficulties. These arise from the time and labor involved both in the construction and in the administration of two complete test forms. When a new selection test is being developed for experimental trial it often seems unduly burdensome to prepare two separate forms of the test merely in order to obtain an estimate of reliability. Furthermore, in a research testing program, time for the administration of an equivalent form of the test is often not conveniently available. In the interests of economy it becomes desirable to set up procedures for extracting an estimate of reliability from a single administration of a single test. One group of such procedures subdivides the total test artificially into two half-length tests and correlates the scores on those. The second group of procedures is based essentially upon the analysis of variance among single test items. The procedures for subdividing the test will be considered in this section, the next section being devoted to procedures based upon analysis of the single items.

#### RELIABILITY OF THE TOTAL TEST FROM PART-TEST CORRELATIONS

If we are to use the correlation between two half-length tests as the basis for estimating the reliability of a full-length test,

some formula must be established expressing the relationship between half-length-test and full-test reliability.

It can readily be shown that the correlation between the sum of any two sets of scores,  $x_1$  and  $x_2$ , and the sum of any two other sets of scores,  $x_3$  and  $x_4$ , is given by the formula

$$r_{(x_1+x_2)(x_3+x_4)} = \frac{r_{13}\sigma_1\sigma_3 + r_{14}\sigma_1\sigma_4 + r_{23}\sigma_2\sigma_3 + r_{24}\sigma_2\sigma_4}{\sqrt{\sigma_1^2 + \sigma_2^2 + 2r_{12}\sigma_1\sigma_2} \sqrt{\sigma_3^2 + \sigma_4^2 + 2r_{34}\sigma_3\sigma_4}}$$

If, now,  $x_1$  and  $x_2$  are scores for two halves of one form of a test and  $x_3$  and  $x_4$  are scores for two halves of another equivalent form of the test, it may seem reasonable to make certain assumptions which considerably simplify the above formula. If we assume that

$$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$$

and

$$r_{12} = r_{13} = r_{14} = r_{23} = r_{24} = r_{34}$$

that is, that the standard deviations of all four distributions of part scores are equal and that the correlations among all four part scores are equal, we get

$$r_{11} = \frac{2r_{\frac{1}{2}\frac{1}{2}}}{1 + r_{\frac{1}{2}\frac{1}{2}}} \quad (3)$$

where  $r_{11}$  is the correlation between two full-length tests and  $r_{\frac{1}{2}\frac{1}{2}}$  is the correlation between two half-length tests. This formula may be generalized to any increase in the length of a test, and it then becomes

$$r_{nn} = \frac{nr_{11}}{1 + (n - 1)r_{11}} \quad (4)$$

where  $r_{nn}$  is the reliability of a test  $n$  times the length of the test from which the observed correlation  $r_{11}$  was obtained.

The assumptions made in arriving at formulas 3 and 4 should be noted. They are (1) equality of standard deviations of the part scores and (2) equality of the part-score intercorrelations. These assumptions are appropriate when (1) the specifications for each part are the same in terms of number of items, dis-



tribution of item difficulties, and distribution of item internal consistency measures (see Chapter 8 on item analysis procedures) and (2) the function tested in later items or trials is not changed as a result of the experience with earlier items or trials. Essentially these same conditions are implied by the assumption of equality of all part-score correlations. That is, the several scores must be truly equivalent in the sense that all parts conform to the same detailed specifications as to sampling of item content, distribution of item difficulty, and internal consistency, and the function measured must not change qualitatively during the course of the test. When these conditions are approximately met, the formula may reasonably be applied.

The reliability of a test may also be estimated from the distribution of differences between scores on the two halves of the test. Under certain simple assumptions the variance of the differences between scores on the two halves of a test is equal to the error variance of scores on the total test. Referring to formula 2 on page 71 and substituting the variance of the half-score differences for the error variance of the test, we have

$$r_{11} = 1 - \frac{\sigma_d^2}{\sigma^2} \quad (5)$$

where  $r_{11}$  is the reliability coefficient for test 1.

$\sigma^2$  is the variance of scores on test 1.

$\sigma_d^2$  is the variance of difference scores between the two halves of test 1.

Formulas 3 and 5 will give identical results when the assumptions for both are completely satisfied, but since they are based upon slightly different assumptions the two may yield slightly different results in individual cases. Formula 5 makes only the assumptions that (1) the variation from true score in a series of measurements of a single individual is constant for all individuals in the group and (2) the errors of measurement in the two half-scores are uncorrelated. When the group tested is fairly homogeneous the first assumption appears quite reasonable. The second assumption is equally involved in any procedure of splitting scores on a single test. The simplicity of formula 5 often provides an attractive computational routine.

## GENERAL EVALUATION OF RELIABILITY ESTIMATION FROM PART SCORES

In general, an estimate of reliability obtained from two parts of the same total test differs from one obtained from the administration of two separate tests in two respects: (1) the two parts are not separately timed and (2) the performances on the two parts are necessarily adjacent or even intermingled in time. A question may be raised as to the comparability of the two part scores, but the same issue arises with any two test scores, whether they stem from artificial subdivisions of the same total test or from separate and distinct test forms.

The extraction of two scores from a test with a single common time limit becomes of critical importance whenever the test is in some degree a speed test. This can be seen most clearly by considering an hypothetical *pure speed test*, in which each individual could do every item if he were given enough time and in which individuals differ only in the number of items which they can do within the limited amount of time available. It is in this case impossible to extract two meaningful scores from a test with a single time limit. The score which an individual makes on a group of items will depend solely on where the items are placed in the test. If two part scores are made up, one of the odd-numbered and the other of the even-numbered items, each individual will *necessarily* have practically identical scores on the two halves, because opportunity to attempt items has been systematically equated for the two half-scores. On the other hand, if one half-score is made up of the first half of the items on the test and the other of the second half, scores on the two halves cannot possibly be compared meaningfully because the individual can score on the second part only when he has completed and achieved a perfect score on the first part.

In practice, no test is an absolutely pure speed test of the sort which we have imagined in the previous paragraph. However, there are many tests that involve speed to a greater or lesser extent. In so far as speed, as distinct from "power" or level of performance with unlimited time, is a factor in the test performances of a group of individuals, the results from split-test procedures for determining reliability will lack meaning. The

amount of distortion of the results will be a function of the extent to which individual differences in score depend upon individual differences in speed of performance.

The second limitation on split-test reliabilities is the lack of time interval between the two performances. In many cases the two performances are not only adjacent in time but even intermingled. This means that the day-to-day fluctuations in conditions (category III) and even the minute-to-minute variance in performance (category IV) tend to be equated for both part scores and to be allocated to systematic rather than error variance. This will make split-test reliabilities in error on the high side, and as far as variance of these types is substantial we may expect a substantial overestimation of the effective reliability of the test.

#### MANNER OF ASSEMBLING PART SCORES

If a test is composed of  $2n$  separate items or parts, there are  $(2n - 1)(2n - 2) \cdots (n + 1)(n - 1) \cdots 1$  ways in which two subtests, each composed of  $n$  items, can be assembled from it. Certain procedures have been proposed for selection from among these many possible alternatives, either on logical grounds or on grounds of convenience. The more usual procedures include:

1. Selecting rationally equivalent groups of items for each half-test.
2. Putting alternate items or trials in each half-test.
3. Putting alternate groups of items or trials in each half-test.
4. Using the first half of the items or trials as one half-test and the second half as the other.

We must now consider the specific merits of these different procedures.

*Rational equivalence as a basis for splitting tests.* Since the total test was presumably constructed to conform to certain specifications (content, difficulty, and the like), it is only reasonable that each of the half-tests should conform to these same specifications. The best guarantee of equivalence in the two half-tests would be to choose equivalent items for each. In other words, the same procedures that were described on page 80 as appropriate in the construction of the two equivalent

forms can be applied to the problem of subdividing the items within a single form. Items selected for each half-test should conform to the specifications for the total test, but within these limits chance should determine which items go in which half of the test.

This appears to be the most defensible procedure for obtaining two half-scores from a test. Its disadvantages are chiefly practical ones: it requires work, and also ideally some data about the individual items, to obtain effective equivalence. It may also involve some sacrifice of convenience in scoring the tests.

*Rationally equivalent and separately timed halves.* A procedure that is a compromise between the use of equivalent forms and the use of halves of a test for estimating reliability is to prepare a test in the form of two rationally equivalent, separately timed halves. These halves can be printed in the same booklet and administered in immediate succession, but each has its own time limits. The scores on the two halves then represent experimentally independent scores, and the speed factor can no longer produce a spurious relationship between the two. At the same time, the two samples of behavior are at least distinct in the segments of time which they occupy.

The preparation of a test in two separately timed halves overcomes most of the practical objections to preparing equivalent forms of a test. No more test items need be assembled than will be used for the experimental form of the test. No extra testing time is required. Some extra work and inconvenience is involved in arranging the test items in equivalent halves, and testing is complicated slightly by the necessity of separate timing for the two halves of the test. However, these are very minor inconveniences. The gain from having two scores which are defined by separate testing operations is quite adequate compensation for the added work in most cases. Where practical considerations rule out the preparation of two complete equivalent forms of a test, this procedure is judged generally to be next in order of desirability.

The additional complications due to separate timing of the halves of a test need not be retained when the test is adopted for routine use in a testing program. It is always possible to combine the two halves into a single test. If this is done, the order of



items will often need to be rearranged, in order to progress from the less to the more difficult items. New norms will be required, of course, for the test with a single time limit.

*Alternate items as a basis for splitting tests.* A procedure which has been widely adopted for splitting the items in a test in order to yield two half-scores is to treat the odd-numbered items as one half-test and the even-numbered items as the other. This procedure has simplicity and objectivity to recommend it. It also is related to the frequent practice of grouping similar items together in a test form and graduating the difficulty of the items from easy to hard. If the items within the test are arranged in this systematic fashion, the odd-even procedure provides a simple way of approximating equivalence in the two half-scores. If there are several successive items on the same topic or of the same type, this procedure automatically divides them evenly between the two half-tests. If the items progress in difficulty, approximate equivalence of the half-tests in difficulty level is guaranteed. However, this reasoning is based on very definite assumptions, which may not be warranted in a given case, as to the manner in which the test items were arranged.

The odd-even items split provides, therefore, only an indirect approach to equivalence in the two half-tests. Particularly with this procedure, certain other issues are raised. If, either because of very close similarity of content or because of moment-to-moment fluctuations in efficiency, performance on successive items tends to be more alike than on items widely separated in the test, we may find that the error of measurement in successive items is not independent but correlated. That is, variance in categories II B and IV B of Table I, which should be treated as error variance, may be common to successive items. By systematically allocating alternate items to the two half-scores, this variance becomes common to both scores and is treated as systematic variance. The odd-even procedure is preeminently the one which makes short-time fluctuations in individual performance operate to give an appearance of reliability rather than one of unreliability.

*Alternate groups of items as a basis for splitting tests.* In some tests, several items may be unduly closely related in content. Examples of this are a group of reading comprehension items all



based on the same passage, a group of items all referring to the same chart or table, or a group of mechanical comprehension items all referring to a single diagram. Such groups of items may share specific content (category II B). If so, it may be preferable to put all the items in a single group into one half-score and to base the two half-scores on alternate groups of items. This procedure reduces somewhat one of the objections to the odd-even items procedure.

*First versus second half as a basis for splitting tests.* In order to avoid the possibility of correlated errors of measurement arising from relatively short-time fluctuations in performance, the procedure has sometimes been adopted of correlating score on the first half of a test with score on the second half. This introduces obvious difficulties whenever the test is not very homogeneous in content. If there is a systematic shift in content or function from the beginning to the end of the test, the first and last half are clearly not equivalent. In the ordinary test of aptitude or achievement, this procedure probably gives a less satisfactory approximation to equivalence in the two half-scores than the odd-even items procedure.

Even when the formal content of a test is homogeneous from beginning to end, as in a series of trials in some complex motor task, the function may change qualitatively for the individual subject. That is, with continued practice the individual may find that the demands and character of the task change. What was initially a problem in discovering correct responses and procedures may, with practice, have changed to a task in developing maximum speed and precision of motor control. Thus, even identity of the external definition and formal requirements of the task cannot guarantee equivalence of the task as faced by the subject.

#### 4. Analysis of variance among items

Any subdivision of a total test into two halves requires a somewhat arbitrary choice from among the very large number of possible ways of making that subdivision. With that in mind, several workers have developed procedures to make use of all the information about consistency of performance from item to item within a test and thus provide a unique estimate of internal

consistency. This type of reliability estimate is analogous to those obtained from subdividing a test and has many of the same characteristics and limitations.

In particular, the procedures for analyzing item consistency are not applicable to a test that involves the element of speed and is administered with a single time limit. The assumption is implicit that the individual has attempted each item. Item characteristics, such as item difficulty, item variance, and item intercorrelations, become quite meaningless when any appreciable fraction of the group has not had time to read and attempt the item.

Moreover, analysis of the items or trials of a test provides an estimate of consistency at a specific time. The temporary factors which were grouped in categories III and IV A of Table I remain relatively constant for each individual during a single test period and are therefore considered as systematic rather than error variance. No estimate of the day-to-day consistency of the individual is possible with those procedures.

#### KUDER-RICHARDSON FORMULAS

An article by Kuder and Richardson<sup>4</sup> presents the development of a series of formulas which extract from the consistency of performance within the items of a test an estimate of reliability. The most general formula assumes only the existence of two hypothetical equivalent tests, equivalence being defined as meaning that for every item  $a, b, c, \dots, n$  in one test there is a comparable item  $A, B, C, \dots, N$  in the other test. To be comparable, a pair of items must have the same difficulty level, the same non-error variance, and the same correlations with the other items in each test. This formula is not soluble, so further assumptions must be made. The first additional assumption is that each of the items in the test measures exactly the same composite of factors as every other, i.e., that the test is perfectly homogeneous.<sup>5</sup> This yields a soluble formula (Kuder-Richardson

<sup>4</sup> G. F. Kuder and M. W. Richardson, "The Theory of Estimation of Test Reliability," *Psychometrika*, 2, 151-160 (1937).

<sup>5</sup> Kuder and Richardson define homogeneity statistically as meaning that the matrix of item correlations is of rank one. This has been shown to be

Formula 8), but one that involves very laborious computation. If a further assumption is made that all the inter-item correlations are equal,<sup>6</sup> a considerably simpler formula results. This formula (Kuder-Richardson Formula 14) is then

$$r_{tt} = \frac{\sigma_t^2 - \Sigma pq}{(\Sigma \sqrt{pq})^2 - \Sigma pq} \cdot \frac{(\Sigma \sqrt{pq})^2}{\sigma_t^2} \quad (6)$$

where  $r_{tt}$  is the estimated reliability.

$\sigma_t^2$  is the variance of the total test.

$p$  is the proportion of individuals passing an item (each item in turn).

$q$  is the proportion of individuals failing an item.

A further simplification of computations may be achieved if it is assumed that each item has the same variance. We then get

$$r_{tt} = \frac{n}{n-1} \cdot \frac{\sigma_t^2 - \Sigma pq}{\sigma_t^2} \quad (7)$$

where  $n$  is the number of items in the test. This is Kuder-Richardson Formula 20 and is the formula which is most likely to be used. An independent derivation of this formula has been presented,<sup>7</sup> which depends upon rather less cramping assumptions than the original ones. This derivation requires the existence of two equivalent tests, where equivalence is defined to mean (1) equal variance for the two tests and (2) equal *average*

untenable when the test contains items varying in difficulty. See R. W. B. Jackson and G. A. Ferguson, *Studies on the Reliability of Tests*, University of Toronto Department of Education Research Bulletin 12, 1941.

<sup>6</sup> This also cannot hold unless the items are of equal difficulty, if one thinks of the item as having a point distribution. That is, if an item is considered as either passed or failed and performance is treated as falling in one or the other of those two categories, the correlation between two items will depend in part on the proportion of individuals succeeding with each. The correlation will be the maximum when these two proportions are equal and will become less as the difference between the two is increased. If in our theoretical discussion, however, we can conceive of each item as corresponding to a continuously distributed underlying variable, this difficulty is perhaps overcome.

<sup>7</sup> See R. W. B. Jackson and G. A. Ferguson, *Studies on the Reliability of Tests*, University of Toronto Department of Education Research Bulletin 12, 1941.

co-variance for the items of each test and of the items of each test with those of the other test. It can be shown that the second half of this definition is in essence a specification of homogeneity of content among the items of the test, but other assumptions are not required. The above formula is quite convenient to compute, requiring only the variance of the test, the number of items, and the difficulties of the separate items. If the assumption of homogeneity of content seems reasonable, it may be considered equivalent to a generalized split-half reliability for the test.

Kuder and Richardson present one additional formula (Formula 21) which reduces computation to the barest minimum. This formula involves the further assumption that all the items are of the same difficulty. The formula can be written

$$r_{tt} = \frac{n}{n-1} \frac{\sigma_t^2 - n\bar{p}\bar{q}}{\sigma_t^2} \quad (8)$$

where  $\bar{p} = M_t/N$ . Thus, this formula requires only the mean and standard deviation of the distribution of test scores and the number of items on the test. However, the assumption of equal difficulty for all items of a test is definitely inaccurate in most cases, and the use of this formula is thereby rendered questionable. It can be shown that any error introduced by this assumption will result in a lowering of the resulting coefficient. This means that Formula 21 represents a minimum estimate of the type of internal consistency that is shown by the relationship between performance on the different test items. It may have some usefulness as such a quick minimum estimate, but it probably cannot be relied upon beyond that point. The general limitations of internal consistency estimates, which are discussed later, apply to this formula.

#### HOYT'S ANALYSIS OF VARIANCE PROCEDURE

The problem of estimating test reliability from consistency of individual performance upon the items of a test has been attacked directly by analysis of variance techniques by Hoyt.<sup>3</sup> Hoyt assumes that the score of an individual on a test may be

<sup>3</sup> C. Hoyt, "Test Reliability Obtained by Analysis of Variance," *Psychometrika*, 6, 153-160 (1941).

divided into four independent (mutually perpendicular) components, as follows:

1. A component common to all individuals and to all items.
2. A component associated with the item.
3. A component associated with the individual.
4. An error component that is independent of (1), (2), and (3).

It is assumed further that the error component of each item is normally distributed, that the variance of the error component is the same for each item, and that the error components for any two distinct items are uncorrelated. When these conditions are met, it is possible to analyze the variance in test scores into the variance contributed by each of the last three components. (The first component is a constant for all items and all individuals and hence is not a source of variance.) Reliability may be estimated from the expression

$$\text{Reliability} = 1 - \frac{\text{Error variance}}{\text{Variance among individuals}} \quad (9)$$

If data are available for a total of  $k$  students on each of  $n$  items, the situation may be illustrated as shown in the table. The  $t$ 's

STUDENT	ITEMS					SCORES
	1	2	3	...	$n$	
1						$t_1$
2						$t_2$
3						$t_3$
$\vdots$						$\vdots$
$k$						$t_k$
TOTALS	$p_1$	$p_2$	$p_3$	...	$p_n$	$\sum_{i=1}^n p_i = \sum_{i=1}^k t_i$

represent scores of individual students, and the  $p$ 's represent numbers of correct responses on the particular items of the test. The sum of squares "among students" is

$$\frac{1}{n} \sum_{i=1}^k t_i^2 - \frac{\left( \sum_{i=1}^k t_i \right)^2}{nk} \quad (10)$$



and the variance among students is this quantity divided by  $k - 1$ . The sum of squares "among items" is

$$\frac{1}{k} \sum_{i=1}^n p_i^2 - \frac{\left( \sum_{i=1}^n p_i \right)^2}{nk} \quad (11)$$

and the variance among items is this quantity divided by  $n - 1$ . The total sum of the squares is

$$\frac{\left( \sum_{i=1}^k t_i \right) \left( nk - \sum_{i=1}^k t_i \right)}{nk} \quad (12)$$

that is, the number of correct responses times the number of incorrect responses divided by the total number of responses. The residual or error sum of squares is the total sum of squares minus the sums of squares attributable to the two systematic factors of individual and item. The error variance is the error sum of squares divided by  $(n - 1)(k - 1)$ .

The results obtained with the above technique can be shown algebraically to be identical with those of Kuder-Richardson Formula 20. It is clear, therefore, that the same restrictions which applied to that formula must apply here also. As we examine the assumptions of this method we note the following points:

1. The test is assumed to be completely homogeneous. No variance is admitted which is common to certain *groups* of items but not to others. Where homogeneity does not obtain, the procedure is not applicable.

2. Independence of error on successive items is assumed. This neglects moment-to-moment variations in efficiency which may effect small groups of items (variance in category IV B).

3. No provision is made for non-attempted items, so that the procedure is not applicable to a speeded test.

4. In addition, it must be remembered that any procedure based on a single testing allocates to systematic variance any day-to-day fluctuations in individual performance.

These considerations apply quite generally to any attempts to estimate reliability from consistency of individual achievement on the items of a single test. The limitations which they impose on the value of this type of estimate are quite severe—so severe that the question has been raised whether this type of estimate should even have the term reliability applied to it. However, we have seen that many of the same objections apply to any split-test procedure based on a single time limit. The chief additional assumption introduced in item consistency procedures is that the test is homogeneous as to the functions measured by each item. Where this assumption is reasonable, the results from item consistency analysis would seem to be equivalent in meaning to those resulting from splitting a test into two parts and to be superior to the latter in that item consistency analysis yields a unique result.

## FACTORS INFLUENCING THE RELIABILITY OF A TEST

The reliability coefficient obtained from administration of a particular form of test to a group of individuals depends on three types of factors. First, we have the series of experimental and statistical operations which were used to define reliability. The various procedures have been considered in the previous section. Second, we have factors relating to the group to whom the test was administered. These factors are, of course, extrinsic to the test itself and serve primarily to confuse the problem of reliability determination. A third group of factors are those that characterize the test itself and the method of its administration.

### *Reliability as a function of group variability*

The feature of a group of individuals that has the most effect on the reliability coefficient obtained for that group (but not on the standard error of measurement) is the range of ability represented therein. It will be remembered that total variance in test score can be divided into error variance and true variance, and that the reliability coefficient may be expressed in the form

$$r = 1 - \frac{\sigma_e^2}{\sigma^2}$$

where  $\sigma_e^2$  represents the error variance and  $\sigma^2$  the total variance. The error variance is that variance which would arise from repeated measurements of the same individual. (For the moment it is assumed that a single value of the error variance characterizes all individuals over quite a range in ability level.) The error variance is not concerned with and does not reflect the amount of variance *between* individuals, as the total variance does. Therefore, the numerator of the expression  $\sigma_e^2/\sigma^2$  may be thought of as essentially constant in groups with different ranges of ability, whereas the denominator increases as the variability of the group increases. Assuming that the standard error of measurement is the same in groups of different ranges of ability, the following equation may be used to express the relationship between reliability coefficients in such groups.

$$\frac{s}{S} = \sqrt{\frac{1 - R_{11}}{1 - r_{11}}} \quad (13)$$

where  $s$  is the standard deviation in the less variable group.

$r_{11}$  is the reliability coefficient in the less variable group.

$S$  is the standard deviation in the more variable group.

$R_{11}$  is the reliability coefficient in the more variable group.

This formula permits the estimation of reliability for a group other than that for which the reliability coefficient was originally computed, if the standard deviation in the new group is known. Sometimes it may be necessary to estimate the reliability in a basic population when one knows the reliability coefficient and standard deviation only for a sample that has been curtailed on some other variable. Thus, data on an achievement test may be available only for that part of a group which had passed a preliminary aptitude measure. One may wish to estimate what the reliability would have been in the total group had no screening been carried out. In this case, the only data available for the *total* group are scores on the screening test. The basic formulas for this general problem of restriction of range are considered in Chapter 6. A simpler variation of these formulas applicable to the reliability coefficient when the equivalent forms correlated have the same standard deviation and the same correlations with

other variables has been reported by Davis.<sup>9</sup> The formula becomes

$$R_{11} = \frac{r_{11} + r_{12} \left( \frac{S_2^2}{s_2^2} - 1 \right)}{1 + r_{12} \left( \frac{S_2^2}{s_2^2} - 1 \right)} \quad (14)$$

where  $r_{11}$  is the reliability coefficient for the curtailed group.

$R_{11}$  is the reliability coefficient for the uncurtailed group.

$r_{12}$  is the correlation between the curtailing variable, 2, and the variable under study in the curtailed group.

$S_2$  is the standard deviation of variable 2 in the uncurtailed group.

$s_2$  is the standard deviation of variable 2 in the curtailed group.

### *Reliability as a function of average ability level*

In the immediately preceding discussion it was assumed that the error variance was the same at different ability levels. This assumption means equal variability in each row or in each column of the bivariate frequency distribution for two forms of the test. The assumption will not necessarily hold in particular cases, especially where the range of abilities under consideration is great. On the one hand, the nature and instructions of a test may cause persons of low ability to rely very heavily upon guessing to determine their answers. Then error variance will be greater at the lower score levels. In another type of situation, the amount done by persons of low ability may be very small compared to the amount for persons of high ability, and as a result the variability may be small because of the restricted range of total scores. Consider a test of speed of reading applied to one group of slow readers able to read an average of 100 words in the unit of time and to another group of fast readers able to read 1000 words in the same unit of time. The variation from test to test in number of words read would seem almost necessarily to be less for the slow readers than for the fast readers,

<sup>9</sup> F. B. Davis, "A Note on Correcting Reliability Coefficients for Range," *J. Educ. Psychol.*, **35**, 500-502 (1944).

and the error variance would therefore be less. Whether the reliability coefficient would be higher for the slow readers would of course depend upon the total variance of scores in the two groups.

Enough has been said to indicate that either the standard error of measurement or the reliability coefficient may vary with the average ability level of the group being studied. The variation with ability level is not statable in general terms and must be determined empirically for any given measuring instrument. The possibility of such a variation should make the test user hesitate to apply data on reliability obtained from a group at one score level to a group at a radically different score level. Reliability data should be based on groups that resemble both in average level and in variability of scores the groups with whom the test is finally to be used.

### *Intrinsic factors affecting reliability*

To the test constructor the important question is: What characteristics of the test itself make for reliability? In responding to this question we must consider (1) the characteristics of the single items, trials, or measurements of which the total score is composed, (2) the number of items, that is, the length of the test, and (3) the conditions under which the items are administered.

The typical test for personnel selection is built up of separately scored items, and the typical score is a linear combination of the separate item scores. Therefore the correlation between groups of items can be expressed in terms of the correlations of each component item with itself and with the other items. Following the standard formula for the correlation of sums, the correlation between one score based on a group of  $a$  items and another score based upon a different group of  $B$  items is

$$r_{(1+2+\dots+a)(1+II+\dots+B)} = \frac{\sum_1^{aB} r_{pQ}}{\sqrt{a + \sum_1^{a^2-a} r_{pq}}} \sqrt{B + \sum_1^{B^2-B} r_{pQ}} \quad (15)$$



That is, the numerator is the sum of all the terms in the matrix  $aB$  of correlations between items in the two parts, and the denominator is the square root of the product of the two matrices  $aa'$  and  $BB'$  of self- and cross-correlations of items within the same part. (The self-correlations are, of course, unity.) When the  $a$  items represent one form (or half) of a test and the  $B$  items represent another equivalent form (or half) of the test, this formula yields the reliability coefficient for the test.

Inspection of this formula shows that in the limiting case where all the item intercorrelations are zero the reliability of the test will also be zero, and at the other limit where the item correlations are all unity the reliability will be unity. In general, the higher the item intercorrelations, the higher the reliability of the test. However, the relationship is a complex one. If paired items or groups of items in the two forms of a test have high correlations, it is possible to obtain a very high reliability coefficient even though the correlations of different items within each test and of non-paired items in the two test forms are quite low. Thus, two forms of a test might be constructed, each containing one-half vocabulary items and one-half block-counting items. Even if the correlation between vocabulary and block-counting items were zero, the reliability of the test might still be high, provided the vocabulary items in one form had a high correlation with the vocabulary items in the other and the block-counting items in one had a high correlation with the block-counting items in the other. We may conclude, therefore, that a high general level of correlation among items of a test is a guarantee of reliability in the test, but that when reliability is defined in terms of equivalent forms low correlation among items does not necessarily mean low reliability.

The size of the item intercorrelations is in part a function of item reliability, in part a function of homogeneity of content from item to item, and in part a function of homogeneity of difficulty level from item to item. Item reliability can be seen in the correlation of pairs of equivalent items from two forms of a test, in so far as it is possible to construct pairs of truly equivalent items. Evidence on homogeneity of content is provided by the correlations among items of a single form of test and by the correlations of non-paired items of different forms. When all

the items of a test appear similar to about the same degree, the distinction between item reliability and item homogeneity disappears. This tends to be the case in tests that are relatively homogeneous in form and content. In such tests, average item intercorrelation gives an expression of the degree of consistency in test performance, this average being a function both of homogeneity of material and consistency of behavior. The dependence of test reliability upon inter-item correlations is seen most clearly in the Kuder-Richardson formulas for estimating reliability (see page 91), which were developed from the formula for the correlation of sums, based on the assumption of complete homogeneity of items. Variations in difficulty level also have a limiting effect on item intercorrelation, since difference in difficulty level between two items sets a ceiling on and generally reduces item intercorrelation.

High item reliability appears unequivocally desirable. That is, if the same functions are measured by two items, the item that measures those functions with greater consistency and precision is to be preferred. A high level of homogeneity of items may, however, be a mixed blessing. Although greater homogeneity generally results in greater reliability, it may do so at the cost of validity. Just as low correlations among tests permit higher multiple correlation from a combination of those tests, so low correlation among items permits a higher validity for the total test composed of those items. Where a single test is used for purposes of practical prediction, a great deal of concern for homogeneity of test materials in order to achieve high reliability is probably a mistake. In analytical studies of human behavior, however, concern for homogeneity of function in the items of a test seems legitimate, and when a test battery is being used increasing the homogeneity of the separate tests in the battery is a legitimate technique for holding down the correlations among different tests. Procedures for selection of items to improve test reliability are discussed further in Chapter 8 on item analysis.

A second major factor in the reliability of a test is the number of items. This is expressed in the generalized Spearman-Brown formula for estimating the reliability of a test of any length  $n$  from a test of unit length (formula 4 given on page 84). How-

ever, this formula assumes that the added items are equal in quality to those of the original test. That is, each added unit of length must show the same variability and same correlations with other units as does the original unit test. For some types of tests the difficulties involved in producing additional large numbers of equally good items must be given serious consideration. It has been shown that the reliability of many existing tests could actually have been increased by omitting the items with the lowest item intercorrelations. Lengthening a test does not guarantee an increase in reliability, if the additional material has lower item intercorrelations than the original. The quality of the items, as discussed in previous paragraphs, may outweigh quantity, and sometimes a brief test built of the most consistent items is preferable to a longer test of less carefully chosen materials.

Finally, reliability is a function of the uniformity of testing conditions. Any variation in testing conditions from one test administration to another may be a source of variance in test performance. This variance must be considered error variance and will have the effect of reducing the reliability of the test. Administrative problems which arise in connection with maintaining standard testing conditions will be discussed further in Chapter 9.

## INTERPRETATION OF ESTIMATES OF RELIABILITY

### *Relative versus absolute measures of precision*

As indicated at the beginning of this chapter, reliability or consistency of performance can be expressed either in absolute or in relative terms. Absolute consistency is indicated by the standard error of measurement. This is the standard error of the distribution of scores all of which are estimates of the same true score. It is given by the formula.

$$s_e = s\sqrt{1 - r_{11}} \quad (16)$$

The standard error of measurement is of particular value when one is interested in applying the information with regard to

consistency to different groups. It has been shown in a previous section that the reliability coefficient depends on the range of ability in the group from which the coefficient was determined. This makes it impossible to apply the coefficient directly to another group differing in variability on the trait in question or to compare directly results from such different groups. The standard error of measurement is usually relatively independent of exact spread of scores. It is reasonable, therefore, to expect the standard error of measurement to remain uniform in groups of approximately the same level of ability. This means that it is possible to apply that value directly to new groups which may differ considerably in variability from the group on which the standard error of measurement was originally determined. Where it is desired to apply reliability data to various different groups, the standard error of measurement has definite advantages.

A second situation in which the standard error of measurement has unique advantages is when a test has been constructed with the specific aim of discriminating at a particular point or within narrow limits of the total range of ability. Thus, a test may have as its purpose the distinguishing of the top 10 per cent of a group, and the items may have been selected with the special purpose of differentiating at this point rather than over the whole range. A test constructed in this way would show several interesting properties. In the first place, the raw score units would not correspond to equal scaled steps of ability at different points on the score scale. The raw score units would correspond to smaller scale units in the neighborhood of the critical score level, so that the test would make finer discriminations of ability in that vicinity. The standard error of measurement in raw score units would be found to increase near the critical level, but if raw scores were converted into scaled scores, scaled in equal increments of ability, it would be found that standard error of measurement in scale units was less in the critical region. The relevant estimate of that test's precision for the special purpose for which it was designed would be the standard error of measurement, in scale units, of individuals scoring at or near the critical score level. General correlational measures of per-



formance over the whole range would provide only an average estimate which would fail to reflect the specific purpose of the test.

There is one further situation in which the standard error of measurement is more useful than a reliability coefficient. This is when we are trying to interpret the score of a particular individual. The standard error of measurement gives us an index of how much variation we may expect in repeated measurements of the same individual and therefore of how accurate we may expect our measurement to be. Knowing the amount of variation expected in repeated measurements, we are better able to judge how much significance to attach to a difference between two individuals or between measures of two traits in the same individual.

For most analytical studies of tests the reliability coefficient is preferred to an absolute measure of consistency such as a standard error of measurement. This is true whenever it is necessary to make comparisons between different tests for the same sample of cases. In general, no two tests have raw scores that are expressed in comparable units of measurement. This makes any direct comparison of absolute measures of consistency from test to test impossible. In such a situation, the only comparison of reliabilities which has meaning is the comparison of measures of relative precision. Consistency of placement within the group, if the group is the same for both tests, provides a meaningful basis for comparing the precision of measurement in two or more different measuring instruments.

Another frequently occurring situation in which the research worker has occasion to use measures of reliability is in the evaluation of the correlations between different tests. The reliability coefficient is the index which can be used in an analytical study and interpretation of the relationships among tests.

### NEED FOR DATA ON RELIABILITY

In general, the importance of information on the reliability of measuring instruments has been adequately recognized in the literature of measurement. If anything, the significance of reliability has been overestimated. It must be remembered that



precision in a measurement procedure is a necessary condition only for obtaining significant relations between different measures and is not an end in itself.

For personnel research the importance of unreliability lies in the effect that unreliability has on the correlation between two different measures. The relationship of the correlation between hypothetical perfectly reliable measures of two variables to the correlation obtained from two sets of actual observations is given by the formula

$$r_{A_{\infty}B_{\infty}} = \frac{r_{AB}}{\sqrt{r_{AA}r_{BB}}} \quad (17)$$

where  $r_{A_{\infty}B_{\infty}}$  is the correlation between perfectly reliable "true" scores on variables  $A$  and  $B$ .

$r_{AB}$  is the correlation of actual scores on  $A$  and  $B$ .

$r_{AA}$  is the reliability of the measure of variable  $A$ .

$r_{BB}$  is the reliability of the measure of variable  $B$ .

This is the formula to estimate true score correlation from the correlation of fallible measures, or to correct correlation coefficients for attenuation due to unreliability of measurement.

Equation 17 can be rewritten in the form

$$r_{AB} = r_{A_{\infty}B_{\infty}} \sqrt{r_{AA}r_{BB}}$$

In this form it is easy to see that the obtained correlation will always be less than the correlation of "true" scores, since the expression under the square root sign must be less than unity. It can also be seen that, as the reliability of either test becomes zero, the obtained correlation between the tests must be zero. Of course, in a limited sample the values of intercorrelations of other variables with a measure having zero reliability will not be *exactly* zero, but the deviations from zero will be no more than chance sampling fluctuations.

When  $r_{A_{\infty}B_{\infty}}$  equals unity in the above expression we find that the maximum value of  $r_{AB}$  is equal to the quantity  $\sqrt{r_{AA}r_{BB}}$ . This quantity gives the ceiling for the possible correlation between two tests of specified reliability and indicates the proportion by which the test intercorrelations are reduced due to errors

of measurement. If we consider one of the two measures,  $B$  for instance, to have perfect reliability, the quantity under the radical becomes  $\sqrt{r_{AA}}$ , and this is the extent to which correlations of variable  $A$  with other variables are reduced due to the unreliability of  $A$  alone.

Reliability is a factor to be considered in variables used both as predictors and as criterion measures. We shall turn our attention to reliability as a factor in a criterion first and then consider reliability as a factor in a prediction test.

### *Reliability data in the evaluation of criteria*

The qualities desired in a criterion measure will be discussed more completely in Chapter 5. We may anticipate at this point, however, by saying that one characteristic desired in a criterion measure is reliability. If we think of the criterion as variable  $A$  in the preceding formulas, we can see that the more reliable the criterion is, the higher the correlation which may theoretically be obtained between the criterion and various predictor variables. If the reliability of the criterion is 0.90, it is theoretically possible to get a correlation of 0.95 between a perfectly reliable test and that criterion. If the reliability of the criterion is 0.64, the theoretical maximum is 0.80; if it is 0.25, the theoretical maximum is 0.50; if it is 0.09 the theoretical maximum is 0.30; if it is 0.00, the theoretical maximum is 0.00. We can see that it is not of *critical* importance that the reliability of a criterion be *high* as long as it is established as definitely greater than zero. Even when the reliability of a criterion is quite low, given that it is definitely greater than zero, it is still possible to obtain fairly substantial correlations between that criterion and reliable tests and to carry out useful statistical analyses in connection with the prediction of that criterion. Given a test or composite of tests with a reliability of 0.90 and a criterion with a reliability of 0.40, it is theoretically possible to obtain a correlation of 0.60 between the two.

When the criterion measure is unreliable, the range of numerical values lying between no relationship and the theoretical maximum relationship is restricted, since the maximum is not unity but  $\sqrt{r_{AA}}$ . The range of values actually obtained for test-criterion correlations will be correspondingly restricted. There-

fore, in order to allow for sampling fluctuations and to get stability in the relative size of the coefficients for different tests, the size of the test population must be increased. If all correlations are reduced by one-half because of attenuation due to an unreliable criterion measure, four times as many cases will be needed in order to obtain the same stability of discrimination between the validity of different tests as would have been required if the criterion had been perfectly reliable. However, this is a practical limitation only, and if the additional cases can be obtained the relative validity of different tests can be determined as accurately for an unreliable as for a reliable criterion. From the validity for the unreliable criterion, it is possible to estimate what the validity would have been if the criterion had been perfectly reliable by the formula

$$r_{A\infty B} = \frac{r_{AB}}{\sqrt{r_{AA}}} \quad (18)$$

provided only that we have a satisfactory estimate of  $r_{AA}$ , the reliability of the criterion.

It is more important that the reliability of a criterion measure be *known* than that it be *high*. This information is needed to establish the fact that the reliability of the criterion is not zero. Unless this can be established with reasonable confidence, further data based upon that criterion measure may be ambiguous. In particular, the finding that none of the experimental tests give appreciable predictions of that criterion will be uninterpretable. We will not know whether the lack of correlation arises because the tests were poorly chosen or because the criterion is essentially unpredictable.

In the AAF air-crew classification program there were a number of cases in which correlations were obtained between aptitude measures and existing records of achievement in training or combat and in which the available tests gave no prediction of those criteria. In many of these cases the nature of the criterion was unfortunately such that no estimate of its reliability could be obtained. For instance, ratings were obtained on combat personnel in certain overseas Air Forces. In these ratings only a single rater evaluated each man, so that no estimate of

consistency in making the ratings was possible. The available classification test data were found to have little or no correlation with these ratings. However, general experience with ratings prepared by a number of relatively untrained raters under conditions of minimum supervision leads one at least to entertain the possibility that the reliability of these ratings was essentially zero and that nothing could possibly have been found that would have correlated with them. In a case such as this, one is in the dilemma of never knowing whether the failure to predict was due to inadequacy of the test battery or to extreme unreliability of the criterion.

A second use for statistics on the reliability of a criterion variable is to indicate the maximum correlation that can possibly be obtained between that criterion and a group of highly reliable tests. If the test composite is assumed to be perfectly reliable, this maximum is  $\sqrt{r_{AA}}$ , where  $r_{AA}$  is the reliability of the criterion. If the reliability of the test composite has been found to be  $r_{BB}$ , then the maximum correlation is  $\sqrt{r_{AA}r_{BB}}$ . The square of this value gives the proportion of variance which test composite and criterion *can possibly have* in common. The square of the correlation between the test composite and the criterion shows the proportion which test composite and criterion *do* have in common. The difference between these two proportions gives an indication of the amount of variance in the criterion which might possibly be predicted by some test battery but is not being predicted by the existing battery. It provides some guide to the probability of significant gains from further research devoted to the prediction of the criterion in question, and consequently some indication as to whether research can still profitably be pursued in that area.

### *Reliability statistics in the analysis of test data*

The qualities which are ultimately desired of a prediction test are that it be valid and that it measure some unique and distinctive quality of the individual. Validity for a test is shown by high correlation between that test and a criterion measure. Uniqueness is shown by low correlation between the test and other tests being used as predictors. The reliability of a pre-

diction test is of interest only as a necessary condition of and as a clue to validity and uniqueness. Information with regard to reliability is of interest during the early stages in the development of a new test because it gives some clue to the possible validity of the test. It is of interest later on in determining to what extent the validity of the test could be increased merely by increasing its reliability. The information can also profitably be studied when the correlations of the test in question with the rest of the tests in the battery are available, in order to determine the uniqueness of each test.

When a new test is being developed, we must be sure that it meets at least minimum standards of reliability. Reliability of the preliminary form should be studied with this in mind. Other things being equal, the more reliable a test is, the higher validity coefficients it may be expected to yield. However, emphasis must be placed on the phrase *other things being equal*. We should rarely sacrifice any feature that promises to contribute to the validity of a test in order to make the test more reliable. Efforts to increase reliability by careful selection and editing of test items are well worth while, but novel and experimental test forms should not be discarded merely because reliability is low by conventional standards.

The reliability of a test can normally be increased by lengthening the test. The increase in reliability is given by formula 4 on page 84. This increase in reliability increases the validity of the lengthened test also. The validity of the lengthened test is given by the formula

$$r_{0(n)} = \frac{r_{01}}{\sqrt{\frac{1}{n} + \left(1 - \frac{1}{n}\right) r_{11}}} \quad (19)$$

where  $r_{0(n)}$  is the validity of the test of length  $n$ .

$r_{01}$  is the validity of the test of unit length.

$r_{11}$  is the reliability of the test of unit length.

Inspection of this formula shows that the validity increases only as a function of the square root of the reliability. The effect of increasing the length of tests of specified reliabilities by various factors is shown in the following table:



$r_{01}$	$r_{11}$	$n$	$r_{0(n)}$
.40	.50	2	.46
.40	.50	5	.52
.40	.50	$\infty$	.57
.40	.80	2	.42
.40	.80	5	.44
.40	.80	$\infty$	.45

It can be seen that, where the test is moderately reliable to begin with, the increments in validity from lengthening it will not be great. In a practical testing situation with limited testing time, the gain from increased reliability must be weighed against the cost in additional testing time. The choice must often be made between lengthening a particular test or adding some additional type of test material. The above formula and illustrations suggest that, once moderate reliability has been achieved, additional testing time will ordinarily be spent more profitably on broadening the scope of measurement by adding new types of test materials. Because adding new tests involves an unspecified loss of time in distributing papers, giving instructions, and the like, it would be difficult to provide an analytical solution of this problem.

A further use of reliability data arises in connection with the analysis of the interrelations among a battery of tests. Our concern here is primarily with evaluating the uniqueness of each test as an independent contributor to the test battery. A test is of value in a battery of tests primarily because it measures functions that are not measured by the other tests in the battery. In analyzing the extent to which a particular test accomplishes this, we must divide the variance in test scores into three fractions: common variance, error variance, and unique variance. The first fraction, the common variance, is that variance which is covered both by the test under study and by other tests in the battery. This variance is predictable from the other tests in the battery, and its amount is given by the square of the multiple correlation between the test in question and all the rest of the tests in the battery. The second fraction of the variance in a test score is the error variance. This is variance that is specific

to a single administration of the particular test, and it cannot be predicted even from the administration of a comparable form of the same test. The amount of variance of this type is determined by the reliability coefficient for the particular test and is equal to  $1 - r_{11}$ . The third fraction of variance for a test, and the fraction with which we are particularly concerned in this discussion, is the fraction which represents genuine, systematic variance in individual behavior (predictable from day to day and from one form to another of the particular test) but which cannot be predicted from the rest of the tests in the battery. This fraction represents the distinctive and unique contribution of the test in question to the total battery. It can be determined only by subtraction, and it is the difference between the reliability coefficient for the test and the squared multiple correlation of the test with the rest of the tests in the battery.

In developing a test battery, particularly one that is to be used for purposes of guidance or for classification into a number of jobs, efficiency demands that each test have a substantial amount of uniqueness. If a particular test score can be predicted from the other tests in the battery, it contributes nothing new of its own and can increase the total predictive power of the battery only through increasing the reliability of the composite score. Only if the test has unique variance can it extend the proportion of the criterion variance which is covered by the battery as a whole. When the uniqueness of each test is the maximum, a limited amount of testing time can cover the maximum scope of human behavior. Any attempt to make an actual determination of the uniqueness of each of the tests in a battery requires information on the reliability of each test.

### SPECIAL PROBLEMS IN RELIABILITY DETERMINATION

The personnel psychologist may encounter several special situations which present their own problems in connection with the determination of reliability. These are situations that render the obtaining of two independent but equivalent measures difficult. A few types will now be discussed, in order to anticipate some of the difficulties which the research worker may encounter.

### *Reliability of speed tests*

It was noted, when the various procedures for estimating reliability were being discussed, that certain sets of operations for estimating reliability became inappropriate for a speeded test. In a speeded test, the amount accomplished within the time limit is a fundamental aspect of the performance. To obtain two independent estimates of the amount that the individual can perform within a limited period of time, two separate segments of time are necessary. No testing procedure based on a single administration with a single time limit can give two independent speed measures. This means that any meaningful reliability estimate for a speeded test must be based on two separate test administrations, either of the same test or of equivalent test forms. Of course, a test may be divided into two separately timed halves and the scores from these used to derive an estimate of the reliability of the test. The separately timed halves become in effect two shorter equivalent tests.

As indicated previously, the role of the speed factor in a test score is a matter of degree. In tests of simple numerical operations or of simple perceptual tasks, it becomes a dominant determiner of scores. It enters to some extent in any test in which each individual does not have time to attempt all the items that he might be able to answer. Whether a particular test is sufficiently speeded to distort reliability estimates based on a single time limit is a matter of judgment. Certainly, the sounder procedure when any doubt arises is to base estimates of reliability on two separately timed segments of testing.

### *Reliability of learning tasks*

In most of the familiar forms of aptitude and achievement tests, it is assumed that the subject does not change appreciably in the function being tested as a result of taking the test. His word knowledge or number skills are considered to be stable functions, not significantly affected by a brief period of testing. In such cases as these, the assumption is probably sound. Some tests, however, are quite clearly learning tasks. This is especially true in apparatus tests of novel motor skills, such as rotary pursuit or two-hand coordination. Trial scores on these tests show a

progressive and quite substantial improvement from trial to trial, so that performance on later trials has been substantially affected by the learning on the earlier trials. Individual subjects not only start on different performance levels but also improve at different rates. The individual's performance is determined not by a single parameter but by two or more.

The problem in estimating the reliability of a learning task arises from the fact that the task is not repeatable, while at the same time the error of measurement in successive trials may not be independent. The task is not repeatable from its very nature as a learning task. Once the individual has learned it or made some progress towards learning it, the task has changed to some extent for him. Though the task is externally the same, it is not the same for the subject to whom it is presented because he has changed. A retest with a second series of trials therefore presents a rather unsatisfactory basis for estimating reliability. The learning changes, which may have affected different individuals in different amounts and different ways, make the two measures non-equivalent. This factor tends to make test-retest correlations with tasks in which fairly rapid learning is taking place underestimates of the correct reliability.

For many learning tasks it is very difficult to devise an equivalent form of the task in order to estimate reliability from the correlation of equivalent forms. These tasks do not represent a sampling from a universe of items. Each one represents a single integrated performance. By what criteria shall we decide that this second complex learning task is or is not equivalent to the first one? Judgment from the superficial appearance of the task will certainly be hazardous. Furthermore, if the tasks are really equivalent, experience with the first may have influenced performance on the second to an unspecified degree.

If one falls back on the procedure of subdividing the period of practice into trials or segments and correlating odd versus even trials or segments, one encounters the other difficulty that was suggested. One may well question whether successive trials or segments represent independent estimates of the individual's ability to do the task, or whether they are not affected by common errors of measurement. That is, adjacent trials or seg-

ments of a learning performance are probably influenced to a certain extent by the same chance combination of temporary conditions. If the adjacent trials are allocated to different half-tests which are then correlated to yield an estimate of reliability, the variance arising from these temporary conditions (categories III and IV in Table I) will be treated as systematic variance. In so far as this is true, the split-test procedures will tend to give an overestimate of reliability.

We have indicated that retest reliability tends to give an underestimate for a learning task, whereas a coefficient based on correlating alternate trials tends to yield an overestimate. If these values are not far apart, either value may be accepted as reasonably accurate. When the difference is substantial, however, an accurate estimate of the reliability of the test can hardly be obtained. To illustrate the amount of the discrepancy, representative reliability coefficients are shown in the table for psychomotor tests used for air-crew classification in the AAF.

TEST	ODD VERSUS EVEN TRIALS	RETEST AFTER 30 DAYS
Two-hand coordination	.80	.87
Finger dexterity	.93	.74
Rudder control	.93	.69
Complex coordination	.91	.83
Rotary pursuit	.94	.74
Discrimination reaction time	.88	.76

$N = 1000$

$N \cong 700$

We have discussed learning tasks in which the individual shows a progressive development of skill. There are also certain tests in which score depends to a very considerable extent on "getting the hang" of the test. This may involve comprehension of rather complex and involved instructions, or hitting on an efficient procedure for carrying out a novel task. In so far as the understanding or technique develops gradually, we encounter the same problems that we discussed in the previous paragraphs on the gradual improvement in skill. If the understanding or technique emerges suddenly as a more or less abrupt insight into the task, the difficulties are of the same type but are exaggerated in their effects.



### *Reliability when the result of performance is known*

A basic assumption in estimating reliability is that the errors of measurement in the two scores being correlated are independent. This means that the performance on the second testing must not be influenced by the result of performance on the first. In the typical printed test, the individual never knows the correctness of his various responses, or at least he does not know those results at the time that he is retested. Sometimes, however, the individual may know how well he did on one testing and that knowledge may be fresh in his mind at the time of the second testing. In certain research tests of pilot proficiency in the AAF, for example, the student carried out a particular maneuver, such as landing the plane, twice in close succession. The two trials were used as scores for estimating reliability. In this situation, the nature of any error made in the first landing attempt was often painfully obvious to the student. Any conscious attempt to avoid the same error on a second trial would naturally lower correspondence between the two performances, and the estimate of reliability of performance would be rendered systematically too low. A similar effect might be found in successive shots at a target, or in any task of coordination in which the amount and character of the error was evident to the subject after each trial. With such tests, it is necessary to have sufficient time elapse between the two trials so that specific memories of previous performance have a chance to fade, and each test becomes experimentally independent of the other.

### *Alternate trials versus alternate practice period as a basis for reliability*

The issue of independence of error arises in yet another context in studies of reliability. There are various types of records, particularly of criterion variables, in which the data are of performance over a number of sessions on different days, but where each day's performance consists of several distinct trials, each scored separately. This was well exemplified in the AAF by the circular error records in bombing training. These records showed error for each bomb dropped. However, several bombs, perhaps five, would be dropped on the same mission under the

same conditions of plane, pilot, bombsight, weather, and other surrounding factors. Reliability for a series such as this could be estimated either by correlating odd- versus even-numbered *trials* (bomb drops in the example cited) through the whole series, or by correlating odd versus even *sessions*, putting all the trials within a single session together into the same score.

We must recognize that a number of sources of variance are substantially constant for all the trials of a given session. In the illustration of bombing error these are such things as pilot, plane, bombsight, weather, altitude, and type of target. These should be considered sources of error variance, since they are not characteristics of the individual whose performance is being evaluated. The trials within a single session have this error variance in common and so do not meet the condition of independence of error. Any procedure that assigns half the trials within a session to one score and half to the other systematically equates this error variance for the two half-scores, and the error variance is made to contribute toward an appearance of reliability. The reliability which results in this way is, of course, spurious. It is generally much better to base reliability estimates on the correlation of alternate *sessions* rather than on correlation of alternate *trials*.

### *Independence as a problem in determining the reliability of ratings and subjective evaluations*

A general problem in estimating the reliability of ratings and similar evaluation procedures is to guarantee the experimental independence of the sets of ratings being correlated. We can recognize two types of interaction between the ratings of two raters. On the one hand, the raters may have collaborated directly in arriving at some of their ratings. On the other hand, each rater's rating may be contaminated by what we may call the local reputation of the man being rated.

In the practical administrative task of evaluating personnel as a part of the routine record system or as a basis for promotion, collaboration is not necessarily bad. It is at least possible that a joint evaluation prepared by two or three men cooperatively is as accurate as the average of the two or three sets of separate ratings. In the evaluation of the reliability of ratings, however,

any collaboration of this type is fatal to obtaining a true estimate. It is quite natural for raters to collaborate—for each to seek the support of another person's opinion and to try to reconcile discrepant opinions. It is imperative in research studies that raters be instructed to make their ratings independently. However, where the rating process has not been closely supervised and where the task has been something of a burden imposed on supervisory personnel, one often suspects that instructions to rate independently were not strictly adhered to.

The more difficult and subtle problem is not that of direct collaboration but rather of indirect contamination by what we may speak of as *local reputation*. In many personnel situations, it is typical for persons in supervisory positions to discuss persons working under them. Through such discussion certain of the individual workers acquire reputations of being good workers, being lazy, being grouchy, and the like. The reputation may, of course, be justified. However, under these circumstances the fact that a man is evaluated consistently by two or more supervisors or co-workers does not necessarily indicate consistency of independent judgments. It may represent consistency of the stereotyped picture of the individual within his group.

Evidence can be cited to support the importance of the factors just discussed. In the AAF, a study was made of the reliability of grades on pilot check rides. For a series of grades on check rides given at a single training school at the level of primary training, the reliability was approximately 0.80. When grades in primary training were correlated with grades in basic training (the second level), which was given at a different station, the intercorrelation was only in the 0.20's. A similar relationship, though not so striking, was found to exist when correlations of certain ratings in a single phase of fighter-pilot operational training were compared with correlations between ratings in two distinct phases. A study of Navy efficiency ratings has shown correlations in the 0.30's between successive ratings of an officer who remained in the same post, but correlations between 0.00 and 0.20 for ratings on shore duty compared with ratings on sea duty.

Special administrative precautions can usually be taken to eliminate direct collaboration between individual raters. There

is probably little that can be done, however, to rule out the subtle effects of general reputation. As long as we must rely upon ratings of two or more raters who operate within the same situation and are acquainted not only with the individual but with his reputation, we must recognize that reliability estimates may represent the generality and consistency of that reputation, as much as or more than consistency in judging the behavior of the individual.

## *The Estimation of Test Validity: Criteria of Proficiency*

Any program of research in personnel selection implies the testing of the selection instruments against some standard of subsequent success on the job in question. A selection program could be set up in terms of professional judgment without including this step of experimental testing. Very possibly, if the personnel psychologist were shrewd, the tests would make a practical contribution to personnel selection. However, one would never *know* whether this was the case or not, and there would be no empirical test of whether one procedure was better or worse than another. A personnel selection program which does not involve empirical checks of the selection procedures against criteria of job success is at best a static and untested one. At worst it may be outright charlatanism.

The most fundamental and most difficult problem in any selection research program is to obtain satisfactory criterion measures of performance on the job, against which to validate selection procedures. This problem is absolutely central, for other research can hardly proceed until a criterion is provided, and the program of research can be only as good as that criterion. Difficulties in test construction and measurement may be very great in certain cases, but these are specific difficulties and work can proceed along different lines even though measurement of certain traits is blocked. Until some solution of the criterion problem has been reached, however, the whole research program is thwarted. Of course, personnel research does not and cannot wait until a completely satisfactory solution of the criterion problem is reached; if it did, no research would be done. However, some compromise solution of the problem of providing a



criterion of success in performance of the job being studied must be reached as a basis for any effective research program in personnel selection procedures.

The naïve worker in personnel selection research turns to the criterion problem as an afterthought. He is likely to assume that routine production records are available in readily usable form, or that he need merely go to instructors or supervisors for usable ratings of the individuals whom he has previously tested. Such faith in the availability of adequate criterion records is usually unjustified. Records which one assumes would be kept are often not kept. Records which a central office has directed to be kept are often either fragmentary or so contaminated by various inaccuracies or ambiguities as to be almost unusable. Some comment has already been made in Chapter 3 on the limitations of ratings as selection devices. These limitations continue to hold, with certain modifications and possibly certain additional difficulties, when ratings are used as criteria of performance on the job. Therefore, almost the first concern of the research worker should be to determine in detail what types of criterion information are actually available, what their characteristics are, and what the possibilities are of supplementing them by gathering additional criterion records specifically for the research project under way. An important goal of the original study and analysis of any job is to determine what criterion measures already exist and to set up hypotheses as to additional ones which seem desirable and possible.

### GENERAL PROBLEMS IN CONNECTION WITH CRITERIA

We have indicated that some solution of the criterion problem must be arrived at before further research can be effective. However, there are all degrees of adequacy of solution of this problem. In any given practical instance a number of possible criterion measures will usually suggest themselves, each of which has some degree of adequacy less than complete, and some degree of practicality. Developing a practical research program involves evaluation of and selection from among these possible criterion measures.

For purposes of discussion we may differentiate three categories of criteria: ultimate, intermediate, and immediate. The ultimate criterion is the complete final goal of a particular type of selection or training. For example, it might have been agreed that the final goal in the selection and training of Air Force bombardiers was that they should under conditions of combat flying drop their bombs in every case with maximum precision upon the designated target. The ultimate goal in the selection and training of insurance salesmen might be that each man sell the maximum amount of insurance which would not be allowed to lapse and that he continue actively as an insurance salesman for an extended period of years. The ultimate criterion for a production line worker might be that he perform his task, maintaining the tempo of the line, with the minimum of defective products requiring rejection upon inspection, that he be personally satisfied with the task to such an extent that he is not a source of unrest and conflict with other workers, and that he continue in the job for an extended period of time. It can be seen that the ultimate goal is stated in very broad terms and in terms that are often not susceptible to practical quantitative evaluation. Furthermore, it is usually not entirely accurate to specify a single and unified ultimate goal. The bombardier had to fire a gun as well as drop bombs. The life insurance salesman must keep records and in many instances manage an office, as well as sell to customers. Even with the production line worker we have indicated considerations of contentment and permanence on the job as well as simple performance. A really complete ultimate criterion is multiple and complex in almost every case. Such a criterion is ultimate in the sense that we cannot look beyond it for any higher or further standard in terms of which to judge the outcomes of a particular personnel program.

In practice, the complete ultimate criterion is rarely, if ever, available for use in psychological research. It may be completely inaccessible, as were most wartime combat criteria needed by psychologists stationed in the continental United States. Even when the complete ultimate criterion is potentially available, waiting for it to mature almost necessarily involves a large time lag. In studying the selection of medical school students, for example, the complete ultimate criterion would be the whole

professional history of each entering student. Because it is embodied in a complex of other interacting factors, the criterion record is difficult to purify and to express in usable quantitative form. Therefore, we are almost always thrown back upon substitute criteria which we judge, either in terms of rational analysis or in terms of empirical evidence, to be related to the ultimate criterion with which we are most fundamentally concerned. These criterion measures we may designate as intermediate or in certain cases immediate criteria.

The term "immediate criterion" is used here merely to differentiate that criterion measure which first becomes available from other partial criteria which become available at various later stages in the course of training or performance of the job in question. For example, in pilot training in the AAF the immediate criterion of flight performance was graduation as opposed to elimination from primary flying school. Intermediate criteria of pilot success included graduation versus elimination in basic, advanced, or transitional training, gunnery scores at the transitional or operational training level, and ratings by supervisory personnel either in advanced training or in the theater of combat operations. In selecting students for engineering school, the immediate criterion might be academic grades in the first semester of the engineering program. Grades throughout later years of the educational program would represent intermediate criteria. The ultimate criterion would have to be sought in some very complex over-all assessment of achievement in professional work over the years. It is worth noting that in the example of the pilot cited above even combat ratings must be considered as intermediate rather than ultimate criteria. We are not ultimately concerned with how a man is *rated* by his superiors but rather with how well he actually *performs* in the crucial situations for which he has been trained.

All immediate and intermediate criteria remain partial, since at best they give only an indication of or approximation to the ultimate goal towards which our selection or training is directed. A program of personnel research must start at an early stage to analyze the available immediate and intermediate criteria in order to appraise the adequacy of each as an approximation to the ultimate criterion.

The ultimate criterion of success in any duty must always be determined on rational grounds. There is no other basis on which this choice can be made. The determination of the ultimate criterion represents an agreement among those who are best qualified to judge as to the objectives of the job, the weight to be attached to each, and the behaviors which represent those objectives. In some cases agreement in selecting the behaviors which define the ultimate criterion may be arrived at quite readily; in most cases the process of defining the ultimate criterion involves prolonged and exacting inquiry.

As one moves from the ultimate towards more and more immediate criteria, there will be more and more room for statistical considerations to supplement the rational in evaluation of the proposed criterion measures. That is, if it is possible to agree on certain behaviors on the final job over a period of time as being the closest approximation to the ultimate goal, it may be possible partially to evaluate the intermediate criteria in terms of empirical data on their relationship to the more nearly ultimate one. For example, it was possible in the AAF to carry out certain studies of the relationship between ratings at successive levels of training. It was also possible to relate certain performance records at successive levels to each other. In evaluating academic records for such a profession as that of accountancy, it might be possible to relate performance in school to subsequent success on and rate of progress through the various examinations leading to the status of Certified Public Accountant.

Statistical studies of partial intermediate criteria in terms of their relationship to more nearly ultimate criterion measures are difficult and time-consuming at best. They often necessitate an extended follow-up of personnel. They are always limited by the adequacy of the more ultimate criteria to which they refer. They are probably best justified in those situations in which a really satisfactory index of the ultimate criterion does exist but in which it can be reached only with great expenditure of time and effort. In those cases, it may be worth while to investigate in an experimental population the relationship of the more accessible and convenient criterion measures to the logically very satisfactory but practically inaccessible criterion.



In practice it often happens that a number of intermediate criteria are available for study, but that no one of these can immediately be identified as approximating the ultimate criterion. Evaluation of the several possible intermediate criteria must then be in considerable measure upon a rational basis. It is necessary to examine each possible criterion measure critically and to judge the relevance of that criterion to the ultimate goal of the selection program. An analysis of the correlations among the several intermediate criteria may be possible, but it may not be clear which measure is being tested and which measure is serving as a standard. Neither measure is ultimate, and it may be that neither has a clearly better rational basis than the other. The investigator is likely to think of one rather than the other measure as being the standard, on the basis of factors such as nearness in time to the ultimate performance, acceptability to supervising personnel, and directness of apparent relationship to the ultimate task. However, the discrimination may not be at all clear-cut in many cases, and the study of interrelationships may be thought of as throwing light on each of the criterion measures. In general, high correlation between different intermediate criterion measures strengthens the rational basis for accepting any one of them as a useful criterion, since each then receives some support from the rational justification of the other. Lack of correlation weakens faith in one or both of the measures, except in so far as each measures distinct aspects of performance for which there is no rational basis to expect intercorrelation.

### EVALUATION OF CRITERION MEASURES

The considerations entering into the evaluation of a criterion measure are analogous to those entering into the selection of a test, though the factors take on a different emphasis. In discussions of tests, it is common to evaluate the test in terms of its validity, its reliability, its objectivity, and its practicality. We can use these same categories in the discussion of criteria, but we may find it desirable to change slightly the flavor of certain of the terms. Validity in a criterion will be thought of less in statistical terms and more in terms of the apparent relevance of the behavior to the ultimate goal of training or operations.



Objectivity is of interest primarily as a condition of freedom from systematic bias in the criterion measure. The meanings of reliability and practicality remain unchanged, but they are probably less important in research on criterion measures than in day-to-day test operations.

### *Validity or relevance*

The quality designated as "relevance to the ultimate goal" is the prime essential of a criterion measure. A criterion measure is relevant as far as the knowledges, skills, and basic aptitudes required for success on it are the same as those required for performance of the ultimate task. That is, the systematic, non-error sources of variance in score on the criterion measure should arise from the same factors that make for ultimate success on the job, combined with the same weights. Theoretically, the relevance of a criterion measure can be determined empirically from its correlation with the ultimate criterion, corrected for the unreliability of both measures. In practice, the *complete* ultimate criterion is never available, and approximations to the ultimate criterion may be extraordinarily difficult to obtain and unsatisfactory on other counts. The result is, as indicated earlier, that the relevance of a particular criterion measure usually must be estimated very largely on rational grounds with only limited help from empirical data. The problem is analogous to that of determining what should be included in an academic achievement test (which is, after all, a type of criterion measure). Rational analysis must be relied on to a very large extent in determining what the goals of instruction are and what content and forms of test item exemplify those goals. In the same way, determining whether a particular type of production record is a relevant indicator of success on the job must depend on a thoughtful analysis of the goals of the job, of the nature of the production record, and of the conditions under which it was obtained. The adequacy of the rational analysis depends on the thoroughness of the analyst's understanding of both the ultimate goals and the immediate criterion measure, and on his basic sagacity.

Relevance is *the* absolutely fundamental requirement of a criterion measure. As far as possible, it is important that *all*

systematic variance in the criterion measure be relevant variance. If the criterion measure possesses any substantial amount of non-chance variance which is irrelevant to (or, worse, negatively related to) the ultimate goal, it is possible that apparently successful selection procedures may be based entirely on this irrelevant variance and have no real validity for the ultimate goals on the job. This is likely to happen whenever performance in a highly academic training situation is used as the criterion of success in some non-academic type of performance. The situation can best be illustrated by an example.

Early in World War II, gunnery schools in the AAF placed a good deal of emphasis in their training program on learning the nomenclature of machine guns and turrets and on being able to express verbally the operation of this equipment. Using grades obtained in this type of program as a criterion, it was possible to develop a battery of verbal tests which gave a substantial prediction of those grades. Actual combat gunnery, however, presented no special verbal demands, and the type of memoriter training referred to above was eventually done away with in favor of more and more training in the actual assembly, maintenance, and firing of the guns. The nature of grades in gunnery school changed correspondingly. With this change in the criterion, the validities of verbal selection tests largely disappeared. Though there is no evidence in this case that the selection of gunners in terms of verbal abilities would have done any actual harm to the final output, it certainly would have been wasteful and ineffective. Selection would have been based on the *irrelevant* variance in the partial criterion of training grades.

There are many other instances, both in and out of the military situation, in which academic grades are used as criterion measures because they are conveniently available and because they appear to possess a rather satisfactory degree of reliability. These grades can generally be predicted with fair success, and the research worker may be lulled into a sense of satisfaction and accomplishment by his success in predicting them. Sometimes, of course, his satisfaction may be justified, because the nature of the training program and of the job may be such that performance in training is based on essentially the same attributes as performance in the job. The possibility is always a real

one, however, that the portion of academic achievement that we predict with our tests is *not* the part that is relevant to later success on the job. Whether the criterion is grades or some other type of record, we must always examine it critically to judge whether the aspects of the criterion which we predict will be relevant for the ultimate goal.

Criterion measures vary widely in the directness and obviousness of their relevance to the ultimate criterion. In some the relevance is so obvious and so universally agreed upon that we may label them *directly relevant* criterion measures. Accuracy of fire by a gunner at an attacking target might be such a directly relevant criterion; so might the amount of insurance sold by an insurance salesman, or the production record of a worker on piece work. If a measure can be found that is by common agreement directly relevant, we turn naturally to it as our first choice in a selection research program. When no agreement on *the* directly relevant criterion measure is possible, we are thrown back on the evaluation of the degree of relevance in less clearly relevant measures.

### Reliability

A necessary but not a sufficient condition for correlation between a criterion measure and the theoretically perfect ultimate criterion of success on a job is that the criterion measure have *some* reliability. That is, the reliability must be greater than zero, because if the reliability of the measure is zero it cannot possibly correlate with *anything*. Scores that reflect purely chance factors, so that they do not even correlate with an equivalent measure, cannot possibly correlate with a measure of any other variable. Evidence with regard to reliability must of necessity be statistical, and this evidence is the first essential in the statistical evaluation of a criterion measure. If data concerning the reliability of a criterion measure do not cause one to reject, with satisfactory confidence, the hypothesis that the reliability of the measure is zero, further use of that measure is futile.

High reliability in a criterion measure is convenient but not critically important. Low reliability in a criterion measure merely attenuates all its relationships with other measures. Low

reliability cannot produce systematic stable relationships of the type that may emerge for a measure low in relevance. Since low reliability is due to random, chance factors, it can only weaken the relationships. The effects of low criterion reliability on test-criterion correlations may be estimated and corrected for by appropriate formulas, as indicated in Chapter 4. However, the values resulting from the correction of initially very low correlation coefficients are subject to a much greater fluctuation from one sample to another than are values obtained from initially reliable scores. This arises from the lower precision of the validity coefficients themselves and also from the sampling errors entering into the estimate of criterion reliability. To compensate for these influences, it is necessary to base validity coefficients for unreliable criteria on substantially larger populations, so that each component statistic is determined with greater precision.

Low reliability is caused by inconsistency of performance in the persons being studied and by fluctuations in the external conditions and in the definition of the task. These may be called intrinsic and extrinsic unreliability, respectively. The influence of intrinsic unreliability can be reduced only by increasing the size and representativeness of the sample of behavior included in the evaluation. One must increase the number of bomb drops, the rounds of ammunition fired, the class sessions observed, the weeks of production record included, and the like. This reduces the proportion of the total variance that is due to chance fluctuations in the individual.

We may usually expect extrinsic unreliability also to be reduced by extending the sample of behavior observed, because chance fluctuations in external conditions should also balance out in the larger sample of behavior. Some compensating factors are introduced here, however. As we extend the sample of behavior observed, there is opportunity for wider fluctuation in external conditions. As the time span for criterion observations is increased, there is more opportunity for variation in weather, seasonal factors, condition of machines and equipment, administrative policy, and the like. In some job situations these factors may be negligible; in others, including aviation, they are not. In these cases, one cannot compensate for unreliability in a criterion by taking larger and larger samples of behavior, because



as the criterion material is increased in quantity it may deteriorate in quality.

The other approach to minimizing extrinsic unreliability lies in control of the external sources of variance. These external factors include those affecting both the performance and the observation of the performance. In military aviation, the conditions influencing the performance were such factors as weather, maintenance and calibration of equipment, competence of other personnel participating in the task, and the exactness of definition of the task. Thus, the accuracy of a particular bomb drop was a function of the turbulence of the air, the condition and calibration of the bombsight, the accuracy in construction of the bomb and its freedom from dents and bent fins, the skill and conscientiousness with which the pilot flew the plane, and circumstances of time of day, altitude, length of bomb run, and so forth, under which the bomb was dropped, as well as the skill with which the bombardier performed his duties. The sources of inconstancy in conditions will be different in any given job and somewhat specific to the particular job, but in general we may expect to find variance arising from equipment, from other participating personnel, and from inequality in the task units which are treated as equal. Reducing variance from these factors, as far as feasible, reduces to such administrative problems as maintenance of equipment, scheduling, and training of the other personnel (such as the pilot referred to above). The degree of control which can be achieved is limited by the extent to which schedules, equipment, and other personnel involved can be kept at a uniform standard under the practical conditions of a large-scale program of training or operations.

The reliability of observation of behavior is influenced by the preciseness of definition of the behavior to be observed, the simplicity of the behavior, the degree to which the behavior is overt, the amount of aid provided by instruments, and the extent of the opportunities for observing the behavior. Efforts to improve the reliability of observation should be directed at these factors. The attention of the observer should be directed toward simpler and more specific components of the behavior which can be observed more adequately at one time. An effort should be made to define the phenomena to be observed in precise terms,



so that all observers will look for the same things. Physical conditions should be arranged so that the observer is advantageously placed to see the behavior on which he must report. Instruments, record forms, and the like should be provided to simplify and standardize the making and recording of observations.

### *Freedom from bias*

Bias in a criterion measure may operate to reduce either its relevance or its reliability or both. The nature of its effect depends on whether the bias happens to cut in a random fashion across the groups being compared or whether it happens to be related to the selection test records in some systematic way. An example may be cited to illustrate this point. In obtaining ratings of military fliers in combat units, there was definite indication that rank and length of service in an organization were positively associated with the rating given. At the same time, standards for personnel selection had been raised during the war, so that the men with less rank and experience tended to have better records on selection tests. The combination of these two historical events led to a situation in which ratings were apparently biased in favor of those with lower test scores. Another type of bias, which probably did not have any systematic effect on relevance but which did lower reliability, was variation in standard from rater to rater. In so far as this source of variation cut at random across all aptitude score levels, it served to attenuate rather than to invalidate the criterion measure.

Bias may occur whenever subgroups of a total population are evaluated in systematically differing ways. The subgroups may represent those taught by a particular instructor or group of instructors, those in a particular school, class, combat theater, factory, or district, those well known as opposed to those only slightly known to rating personnel, and the like. Bias may arise within subjective evaluation standards, as when one rater is more lenient in his interpretation of a rating scale than another. It may also arise within external conditions, as when enemy opposition is much more severe in one combat theater than in another or when the atmosphere of the territory in which one group of soap salesmen must operate holds much less soot and dirt than the atmosphere in another district. Bias is a much more serious

matter than low reliability because it may affect systematically the comparisons in which we are interested and thus produce spurious results. Bias is not so universally undesirable as low relevance, since the data may often be gathered in such a way that the source of bias is randomized with regard to the factor being studied.

Subjective sources of bias may be reduced somewhat by the same steps used to obtain more reliable observation. External sources of bias may sometimes be allowed for statistically, if they are sufficiently constant and lawful in nature so that they may be recognized. Thus, in a criterion measure for soap salesmen, a correction may be made for the sootiness of the territory in which a given man works. In evaluating performance of insurance salesmen, wealth of the region and other factors influencing the total amount of insurance sold would have to be considered. Any criterion measure should be studied to determine the possible biasing factors and to correct for these or at least see that they are randomized with respect to the test variable being studied.

### *Practicality*

Considerations of convenience and economy perhaps do not bear the same weight in a criterion measure that is to be used for a special research project that they do in a test that is to be used routinely from day to day. However, there are very real limits to the amount of labor that may be undertaken or the amount of disturbance of routine procedures that will be tolerated in the gathering of criterion records. Unless the project has been well sold to operating personnel, any program for gathering criterion information which calls for additional work on their part may well be resented. Even when the value and significance of the project have been demonstrated effectively, the good will which has been established can soon be exhausted by a program that requires laborious assembly of ratings or bothersome interference with the routine of school or factory, unless the group concerned can see in the criterion measures some value that is much more immediate than a relatively remote and nebulous personnel research program. Realistic evaluation

of the burden which any program for assembling criterion data represents is a necessity for the research worker.

### TYPES OF CRITERION MEASURES

The criterion measures which may be found or developed in a training or job situation are of two broad types. One type consists of procedures for evaluating performance in a specific defined task. This type of criterion measure is exemplified by an objective achievement examination, a work sample test in which the subject is required to produce a sample of the type of work done on the job, or a rating of the subject's behavior in a specific situation, such as the teaching of a single class period or the conduct of a particular sales interview. All these have the common characteristic of being based on a single limited segment of behavior. Some of them have in addition the characteristic of being based on a quite uniform and closely controlled situation. In contrast, other criterion measures represent general summary evaluations of a total phase or large unit of training or on-the-job performance. These are exemplified by average grades for a semester of college work, by periodic supervisory efficiency ratings, by piece-work records over a period of time, and by promotions and other evidences of achievement.

Each of these two types of criterion has its advantages and its limitations, and each can claim a place in a program of criterion development. Specific evaluations of a limited unit of performance have the great advantage that they make possible a relatively exact statement and a relatively high degree of control of the criterion situation and of the conditions of observation of the behavior in which the research worker is interested. This same degree of specification and control of the content and conditions of the evaluation can probably never be achieved in a summary evaluation of an extended period of training or work. On the other hand, the summary evaluation covers a wider scope, in terms of amount and variety of behavior, which cannot be compressed into a limited test or observation period.

When a job is studied by a personnel psychologist, he often finds that certain criterion measures are already available in existing records. These are generally the type referred to as

summary evaluations. As the research worker becomes increasingly well acquainted with the job, he is likely to become sensitive to the limitations of these existing summary records and to see a number of possibilities for developing specific procedures for testing and observing precise units of performance. The development of these procedures represents one major contribution of the personnel psychologist. At the same time he must assume a responsibility for studying, improving, and supplementing the summary evaluations that already exist.

*Specific criteria of performance in a limited behavior unit*

The specific evaluation of behavior within a limited unit may be subdivided into evaluations of knowledge and information about the duties of a job and appraisals of actual performance of the job. Evaluations of knowledge and information about a job take the form of the traditional test, and present the same set of problems that are encountered in any achievement test of problems that are encountered in any achievement test of construction. Measures of performance may be further subdivided, depending on whether appraisal of the performance is based on an objective instrumental record of the performance, on subjective scoring of items of the performance, or on subjective rating of the performance as a whole. These three categories are not entirely separate and discrete, but they represent identifiable points on a recording continuum from objectivity to subjectivity.

At one extreme of the objective-subjective dimension with respect to criterion performance measures, the behavior itself yields a persisting record, and the observer enters only to transcribe or score the record. Since the record persists, any necessary amount of time can be devoted to scoring and any necessary repetition or revision of the scoring can be carried out to make sure that the inaccuracy or bias of the observer is reduced to an insignificant amount. One type of such objective record is, of course, the objective examination. However, this is usually limited to knowledge about, rather than performance of, job duties. In some cases, a written examination may approach quite closely the actual functions performed on the job. In the Air Force it was possible by written examinations to come closer to essential features of the work of the navigator than that of

either the pilot or the bombardier. In such fields as accountancy or bookkeeping one may anticipate that a well-constructed written examination would come close to assessing a number of features of the performance on the job itself. In other tasks, such as selling automobiles or operating a milling machine, a written examination has very little in common with the job situation. In addition to written tests, one can conceive of a number of situations in which a persisting record of the performance of the individual is possible. Gun-camera records of the point of aim, relative to the attacking fighter plane, provided permanent criterion records in aerial gunnery. Various instrumental devices have been proposed to record the slip or skid of a plane in a turn or to indicate the amount of control movement that the pilot found it necessary to use. Any task in which a worker is required to produce a standard sample product has this character. In almost any field, certain permanent objective records of behavior suggest themselves.

The middle point on the scale, subjective scoring of items of performance, is encountered whenever the behavior leaves no lasting record but when specific segments of behavior may be evaluated as they occur in terms of relatively simple and analytical judgments. In the AAF an experimental pilot flight check was developed which depended very heavily on observations of change in the reading of an altimeter, estimates of angle of the plane in a turn, position on the landing field of initial contact in a landing, amount of bounce in landing, and the like. For certain air-crew duties check lists were developed on which an observer checked the occurrence or non-occurrence of each of the elements in a standard pre-flight check. Similar check-list procedures were developed for evaluating assembly of the caliber .50 machine gun and for other tasks which the gunner was expected to carry out under operational conditions. Analogous procedures can be applied to the observation of any job performance that is sufficiently standardized so that it can be broken down into a number of generally accepted units or elements of good performance. Thus, it might be possible to set up a check list for the essential steps taken by a milling machine operator in setting up work in his machine or by a weaver in setting up a loom, to observe workers in a test situation, and to score them



on the completeness with which they carried out the necessary steps in the necessary sequence.

This type of criterion measure extends over quite a range of the objective-subjective dimension, depending on the degree to which the performance can be analyzed into simple and discrete components. That is, we may find at one extreme a task in which the successive elements are each quite separate and follow one another in a necessary sequence. We may find another task in which the units are more complex and the order less definitely established. We may find one task in which the observation is of so simple a thing as the movement of an instrument needle on its dial, while in another task the judgment may be much more complex and may involve such a decision as whether the worker did or did not inspect his machine properly before starting work. In general, we may anticipate that the observer will enter as a more and more significant factor in the final result as the complexity of the components to be observed increases.

At the other end of the objective-subjective dimension we encounter the relatively unanalytical rating of a complete sequence of behavior. This is well illustrated by the grade recorded on the standard check flight used by the AAF during the war to evaluate progress of a student in pilot training. This grade represented the synthetic evaluation of the complete segment of behavior occurring in a ride of perhaps an hour. It was an unanalyzed clinical judgment in which it was no longer possible to identify specific behavior units and for which it was impossible to determine the way in which the items of behavior were weighted in the final composite judgment. Other examples of this end of the dimension include ratings of a period of practice teaching, ratings of a sales interview, and the like.

To recapitulate, we find a variety of measures of behavior in a test situation. These differ in the degree to which the evaluation is mediated by the observer. At one extreme, the behavior itself leaves a permanent record which may be scored or evaluated at leisure, making the evaluation both easier and repeatable. Such an evaluation can to a very large extent be freed of the influence of the observer. Toward the middle of the scale are encountered situations in which the behavior leaves no lasting record but may still be analyzed into rather simple units. Here

the necessary observation can be defined in terms of observable readings of instruments, occurrence or non-occurrence of simple items of behavior, and the like, which require only relatively direct perception on the part of the observer. These are mediated by the observer in the sense that his on-the-spot evaluation of the behavior provides the only available record, but the judgments can be defined in such simple and definite terms that we may anticipate that individual standards of judgment will be of minor importance. As the situation becomes more complex and the required observation more difficult to define exactly, we may anticipate that the mediation by the observer will become more and more important, until it reaches a maximum in the undefined rating of a complete behavior sequence.

The reader will recognize in this discussion an elaboration of the conditions fostering objective evaluation at one extreme and subjective evaluation at the other. The important points to remember are that objectivity-subjectivity is a continuum and that the conditions making for objectivity are persistence of the trace of the behavior and simplicity and precise definition of the phenomena to be observed.

Four types of criterion measures based upon a specific limited segment of behavior have been identified in the previous discussion. These are (1) tests of knowledge and information, (2) objective performance scores, (3) observer scored job samples, and (4) rated job samples. We shall now consider each of these, indicating its advantages and also its particular limitations. The illustrative material will be based very largely on the experiences with criterion measures for air-crew performance in the Army Air Forces.

#### EVALUATION OF KNOWLEDGE AND UNDERSTANDING

Knowledge and understanding of the duties of a job are probably the aspects of proficiency that are most readily evaluated. They have the great advantage of fitting naturally into the form of the objective achievement examination. Examinations of this type having quite adequate reliability can usually be developed for use as criterion measures. This type of examination is also highly satisfactory from the standpoint of freedom from bias and of administrative convenience and practicality. As

has been indicated above, it is on the count of relevance to the ultimate criterion that its value most often seems limited. The gap between knowledge about a job and the ultimate criterion of performance on the job often seems very large. This will differ with the type of job, of course, and the more the particular job involves abstract and intellectual duties the more plausible it is to evaluate proficiency in terms of a printed achievement test.

In the preparation of a printed proficiency test of knowledge about a job, the essential task is to integrate the specialized test construction skills of the personnel psychologist with the knowledge of the job possessed by experienced job personnel. That is, the outline for the test in terms of topics covered and amount of emphasis given to each must seem appropriate to those who know well the work of the job itself. At the same time, the specific elements of knowledge included in test items must be those which the job specialists agree to be important. Likewise the accuracy, up-to-dateness, and keying of test items must be certified by those experienced supervisory personnel in the job situation who are in a position to know the currently accepted job practices and procedures. The personnel psychologist has the responsibility of integrating the suggestions and criticisms from these sources and of translating them into technically adequate test items.

#### OBJECTIVE PERFORMANCE SCORES

To the scientifically trained research worker, the objective performance score promises to be the ideal criterion for much of selection and training research. Since the ultimate criterion of job success is almost always a performance of some sort, records from appropriate types of performance under experimental conditions seem likely to be the most directly relevant to that ultimate criterion. At the same time, the fact that the task leaves a permanent and objective record minimizes the possibility of observer unreliability or observer bias entering in to attenuate or prejudice the conclusions. There are, however, certain qualifications to this so far very attractive picture. These qualifications concern both the feasibility of obtaining an objective record of the types of behavior that are significant in many

jobs and the problem of specifying and controlling the external conditions of the task so that it is possible to present a standard task to any subject at any time and at any place.

Because of the attractive possibilities which this type of measure presented, objective performance records were sought by the AAF Aviation Psychology Program for every air-crew specialty with which it was concerned. This involved setting up for research purposes special performance situations which yielded a direct record, at the same time fully exploring and exploiting existing performance records of the training program. Experimental situations involved in some cases actual in-flight performance; in other cases they involved ground tests or trainers. The flight situation appeared to involve most directly and completely the type of performance for which the individual was being trained. On the other hand, the actual conditions of flight added a number of additional complications to the problem of controlling the test situation. The types of objective records which may be obtained for use as criteria can be illustrated by three examples of objective performance scores which were set up for research purposes.

The first example is a criterion of skill in flexible gunnery. The flexible gunner in a bombing plane controlled a gun or gun turret and fired at attacking enemy planes. The accuracy of his fire under combat conditions represented one major feature of the ultimate criterion of gunnery performance. To approximate this, records of firing were obtained on training flights, with a gun camera attached to combat-type equipment. The gun camera took motion pictures of the point at which the gun was aimed and provided an objective record of the gunner's performance. Subsequent measurement of the point of aim provided a score indicating the accuracy with which each gunner followed the attacking target. Each gunner was tested with a series of fighter attacks. The pilots of both the bomber and fighter planes had been instructed on the standard manner in which these attacks were to be flown, in order to make the task as uniform as possible from man to man.

As an investigation of the course of learning and the amount of practice needed in bombing training, a project was undertaken involving bombing under specified experimental conditions. The

permanent record was the photograph of the bomb strike on a standard desert target. The experiment was devised with controls on the factors of airplane, pilot, bombsight, bombing altitude, length of bomb run, and a number of other variables. The photographs could subsequently be measured to determine distance of each bomb drop from the target.

A rather different type of permanent objective record was obtained for navigators by using the logs of a series of formation navigation missions. The observations and calculations in a navigator's flight log provide an objective statement of where he believed himself to be at specified times during the trip. For the log to become an objective performance measure it was necessary only that the conditions of flight be uniform for different navigators and that there be some way of knowing the *actual* flight course of each plane. The flight for all men in the group was standardized by having the group fly each mission together in formation. The actual flight course was determined by specially trained graduate navigators in the leading plane of the formation, who provided standard values against which the logs of the student navigators could be checked. Additional precautions were taken in order to obtain the maximum amount of standardization within each flight and from flight to flight. The criterion score for each navigator was based on the discrepancies between his log and the official values determined for the formation.

A number of situations in which a comparable type of performance record could be obtained in industry will readily occur to the reader. For example, typists are often evaluated by a standard work sample of typing, graded for speed and accuracy. In evaluating machinists, each man could be required to produce a standard product, and his work could be evaluated by the accuracy with which the specified dimensions were met and the speed with which the task was finished. Accountants could be given a standard set of books and business records to examine and be evaluated on the speed and accuracy with which they analyzed those materials.

It has been indicated that complete specification and rigid control of the external conditions are necessary if objective performance records are to fulfill their promise as the ideal criterion



of proficiency in job duties. It is at this point that one of the major limitations of this type of criterion measure is found. It is fundamentally extremely difficult to obtain the desired degree of specification and control. If we consider the aviation examples just described, we see that the conditions surrounding the performance of an air-crew member in the air are enormously complex. They involve, first, all the conditions of temperature, visibility, and turbulence which constitute the weather. Only a limited control is possible for these factors, by restricting times of day during which flights are flown or by canceling flights when weather conditions are too unfavorable. Under the practical pressures of a military time schedule even this amount of control was not always possible.

The second major group of factors requiring control deals with equipment. Calibration and maintenance of the bomb-sights, accuracy of alignment of the drift meters, uniformity in the compasses and air speed meters, and so on, for the many guides on which the bombardier, navigator, or pilot must rely for information, all influence an individual's score. When the typical personnel error has been reduced to quite a small size, a small instrument or equipment error may become a major factor in determining the final result. Just chance variation in the physical characteristics among practice bombs, for example, might constitute a substantial part of the error in good bombing. Theoretically, these factors can be reduced to minor importance by perfect maintenance of equipment. However, aviation research in wartime was done not under theoretical but under actual conditions of maintenance, and this will be true for any practical program of military or industrial personnel research. Under these conditions, lack of uniformity of equipment is likely to constitute a major problem in the use of objective records as criterion measures.

The third type of factor which often complicates the evaluation of objective performance scores is the influence of persons other than the individual being evaluated. In all the examples of air-crew performances which were cited the test situation depended not only on the equipment being used but on the individuals using it. Planes for bombardier and navigator training were flown by pilots who exhibited varying degrees of skill and varying

degrees of conscientiousness in following the detailed instructions which were given to them. In some cases still other persons had a role in defining the task presented to the person being tested. In spite of the best efforts of research and supervisory personnel, standardization among these various individuals who entered into the test situation, and to whom the test situation was just another job, was only moderately good.

The net result of the operation of such factors is that the reliability of objective scores obtained from an experimental session in a complex job situation may be quite low. This was true for certain of the air-crew measures which have been described. Thus, navigation formation flights yielded correlations between scores for a man on two missions ranging from  $-0.16$  to  $0.27$ , with a median value for 25 coefficients of  $0.02$ . A somewhat more favorable picture is given by gun camera records for flexible gunners when these were obtained under carefully controlled conditions. An examination of 18 coefficients shows a range from  $-0.11$  to  $0.75$  ( $N = 16$  to  $24$ ), with a median value of  $0.42$ . Even this value is none too high. The problem of reliability appears to be quite a critical one for this type of criterion.

#### OBSERVER-SCORED JOB SAMPLES

This category is used to cover those instances in which the performance itself leaves no lasting record, so that it must be evaluated as it occurs, but in which the evaluation can take the form of direct observation and scoring of limited and rather well-defined units of behavior. The types of behavior range from those in which a recording instrument could readily be substituted for the observer, if an instrument happened to be conveniently available, to those that require a moderate amount of synthesis and interpretation by the observer. A number of procedures of this type were developed in the AAF Aviation Psychology Program. There were many samples of behavior which left no permanent record of performance, and these procedures seemed to provide the nearest approach to the objectivity of such a record. They were variously designated as "phase checks," "performance checks," "objective scales of flying skill," and the like.

One group of the measures developed in the AAF consisted of an experimental series of scales of flying skill developed by psychologists working in problems of pilot training. In these, certain standard maneuvers were specified, to be flown in a defined sequence. A score card was prepared which specified the aspects of each maneuver which were to be observed and scored by the check pilot as the maneuver was performed by the student. Thus, on a steep turn the observer might have to score the angle of bank, the time to complete the turn, and the change of altitude during the turn. On a landing the observer might score the part of the field landed in, the amount of bounce in the landing, and the attitude of the plane at the time of landing. In developing the maneuvers and the scoring procedures, attention was directed toward making all observations as simple and as quantitative as possible. Many of them were expressed in such quantitative terms as feet of altitude, degrees of heading, or miles per hour of air speed. These were supplemented where necessary by more qualitative judgments of coordination of controls, amount of bounce, and so forth.

A somewhat different type of scored job sample was the "phase check" or "performance check." This type was developed most extensively in connection with the gunnery training program. A typical check was that on the performance of stripping and assembling the caliber .50 machine gun. The task was analyzed into a sequence of component operations. A score sheet was prepared listing each step in the sequence. The observer checked the student step by step on the score sheet, indicating by a simple check mark whether he did or did not perform each required step adequately and at the proper point in the sequence. This type of check is well adapted to tasks for which a standard sequence of operations is required and to tasks that are readily analyzed into a number of component elements.

When the routines are less readily specified or the operations are more complexly integrated, the checking procedure becomes more difficult, more subjective, and perhaps less satisfactory. However, in the AAF a number of check procedures were developed for quite complex sequences of tasks. Thus, a flight check for bombardiers was developed which covered the complete course of a practice bombing mission, including pre-flight

inspections and all flight operations up to and including the simulated bombing of the target. A similar check was developed for radar observers. Very little evidence was actually obtained as to the objectivity of application of complex flight checks such as these. It seemed clear, however, that they required a checker with a high level of experience and competence in the job duties, and that some special training in the use of the check was needed.

Various analogous uses of observer-scored performance tests and of check procedures in industrial situations suggest themselves. In evaluating the proficiency of truck or bus drivers, a standard sequence of maneuvers with the vehicle could readily be set up. Backing and turning with limited tolerances for error could be observed upon a test route and scored by the observer. Smoothness of starting, shifting gears, starting on a hill, and the like could be observed. From the observation of performance in a number of specific test situations, a composite score could be derived for the test as a whole. In other cases, in which a job or an aspect of a job could be broken up into observable fractions, the phase check type of procedure may prove applicable. Thus, it might prove feasible to analyze the operations involved in setting up work on a lathe into a sequence of necessary steps and to develop a score card for checking the worker's adequacy in carrying through each of the successive steps. A similar type of check procedure might be developed for the apprentice weaver who was learning to set up a loom.

In potential relevance to the ultimate job criterion, observer-scored job samples such as these are second only to the direct objective record of performance. Their directness of relevance is somewhat less in so far as an observer is introduced between the behavior of the subject and the record of that behavior. However, the use of an observer permits flexibility and scope considerably greater than are possible for an objective record. The opportunities for bias vary greatly for different instruments within this category, depending on the type of observation required of the observer. If the observer serves as a recorder of instrument readings or of simple precisely defined behavior items, bias is minimized. As more interpretation by the observer is required or permitted, more variation from observer to ob-

server, from time to time, and from place to place may be expected.

In measures of the type now being discussed, reliability continues to be a critical problem. The observations are subject to much the same disturbing factors as those discussed in connection with performance records. However, with the observer taking a significant role in the evaluation, the factors are somewhat changed. On the credit side, the interpretation of the observer makes possible some allowance for variation in the objective external conditions. Thus, in aerial checks the observer can make some allowance for conditions of visibility, turbulence, and the like in evaluating the performance of a particular individual at a particular time. Another advantage is the increased range of situations that can be evaluated if the observer is included. This increased range permits broadening the base of the observations, and the additional types of data included may be expected to contribute both to the reliability and to the relevance of the total score. On the debit side are, of course, the fluctuations of the observer from moment to moment and variations from observer to observer in standards of evaluation. Effective use of checks of this type would require the scope of the evaluation to be increased as much as possible, and at the same time the points observed to be so completely defined that variations from observer to observer are minimized.

Data from AAF studies on the reliability of observer-scored job samples are rather limited, but certain illustrative figures may be cited. For a series of elementary pilot training maneuvers, a median day-to-day retest reliability of 0.08 was obtained for 18 separate single items of performance scored on successive days by different check pilots. On a different specific group of items the reliability of a total scale of 16 items, selected from a larger group of 24 items in terms of ability to discriminate between men with differing amounts of training, was 0.50 for 41 men with 55 hours of training and 0.39 for 35 men with 15 hours of training. A complete scale for basic instrument flying gave a test-retest reliability of 0.43 for 55 cases. This was, again, for retest on different days with different check pilots. These reliabilities are for procedures upon which a good deal of



developmental work had been done, so that it is clear that reliability remains a problem for materials of this type.

#### RATED JOB SAMPLES

Although the objectivity of actual performance measures in the job situation makes them appear particularly attractive as types of criterion record, many aspects of performance on a job do not leave any persisting record which can subsequently be studied. Some of these, as we have seen, can be analyzed into rather specific items of behavior, and the adequacy with which the individual carries out the work of the job can be checked by observing these specific behavior sequences. But a wide range of job functions still remains which cannot be analyzed in this fashion or which are extremely resistant to analysis. In many cases it does not seem possible to break up the job performance into specific elements or items which can be observed separately. Thus, if we wish to observe the performance of a teacher during a period of practice teaching and to evaluate that performance, it is probably not feasible to set up a list of prescribed behaviors and check whether the teacher does or does not perform each one of the prescribed items. The situation is sufficiently fluid so that we probably cannot be sure in advance just which behaviors the teacher will be called upon to display. At best we are probably limited to analyzing teaching performance into certain major aspects and arriving at a rating of the performance of the teacher with reference to each of those aspects. These ratings then refer necessarily to complex phases of behavior occurring in a complex behavior sequence, and the rating necessarily represents a subjective evaluation of each.

This type of rating is well illustrated by the pilot check flight which was in regular use in the AAF during World War II as a procedure for evaluating the performance of student pilots in training. In this evaluation, the student flew with a check pilot. He went through a series of maneuvers appropriate to his level of training, the particular choice and sequence of maneuvers being determined by the check pilot. After a flight of varying duration, for which an hour might be roughly a representative figure, the check pilot recorded the grade for the flight, together with such comments as he considered appropriate. The final

grade represented a complex clinical evaluation of the performance of the student during that flight, modified to some extent by the observer's previous acquaintance with the student or the student's record.

The subjectivity of the above type of procedure and its extreme dependence on the standards and judgments of the observer are obvious. Some degree of standardization may be achieved by centralized training of instructors and check pilots, by review of the ratings given by individual check pilots, by standardization boards, and the like, but even so individual standards may be expected to vary significantly. Unless a strenuous effort at standardization is made, variation from observer to observer is likely to become enormous. As a compensating advantage this procedure makes use of a synthetic judgment of total performance. There *may* be aspects of flight performance which are lost in an analytical approach, so that a scoring of elementary items of performance can never give an entirely adequate evaluation of the over-all quality of flying. As far as that is true the synthetic rating has advantages. It can be argued further that the rating procedures make it possible to allow for the external conditions under which the flight was made. However the essential subjectivity of these ratings leaves them always suspect as criterion measures.

It would be possible to use ratings of a specific sample of job performance for any type of a job that permitted observation of a specified unit of activity on the job. As suggested above, rating procedures may be used in evaluating performance during a session of practice teaching. Again, it would be possible to observe a sales interview and rate the performance of the salesman at that time. In the records of at least one civilian air line, ratings of the co-pilot have been systematically reported by the first pilot of the plane at the end of each trip. These ratings refer to various aspects of the performance of the co-pilot during the trip in question. In many other cases similar opportunities for observing a specific segment of job performance present themselves, and ratings of such performances are generally possible.

Some general discussion of the problems involved in using ratings has already been presented in Chapter 3. We should

probably feel that ratings more likely provide valid indications of individual ability in those jobs in which the ratings are based on the observation of a specific sample of behavior. Here we at least have some concrete behavior to serve as a basis for judgment of the individual. This is certainly a vast improvement over ratings made after the fact and based on general impression rather than on any specific samples of observed behavior. The pilot who has just flown an eight-hour mission with a co-pilot, during which time he has called on that co-pilot to assist him in various ways and to perform various specific duties, is in about as good a position to evaluate that co-pilot with regard to the performance of those duties as he could ever be. However, even in this situation the problems of subjectivity in the use of ratings remain very real ones.

In general one would be inclined to conclude that rating procedures are inferior to those types of procedures previously discussed in relevance to the ultimate criterion and particularly in freedom from bias. The lowered relevance arises from the fact that a stage of interpretation comes between what the man did and the score he receives. This interpretation imposes one more step between the performance and the ultimate criterion. Though we may agree that a plane commander's success in maintaining the morale of his crew, for example, is an index of how competent he will be in his combat duties, we would be less willing to grant that how he appears to an observer to motivate his crew is such an index. The additional step of interpretation seems inevitably to weaken the rationale for the evaluation procedure.

It is especially in the matter of bias that rating procedures appear to be weak. In these procedures to a very large extent each rater provides his own standard. This standard varies from rater to rater, from time to time, and from place to place. Comparison of groups in different schools or factories becomes meaningless because of local variations in standards of evaluation, so that large-scale and long-time studies become impossible. Bias becomes an important matter, especially when two distinct groups are compared. Even when systematic procedures are introduced to assign to each rater members of each of the experimental groups which are being compared, bias is still pos-

sible. If the raters are aware of the group to which a particular individual belongs and if they are prejudiced in favor of some one of the particular training programs that are being compared, it is entirely possible for the rating of a man to reflect a bias towards the group of which he is a member. Therefore, especially in investigations of particular experimental or training procedures the use of ratings as a criterion must be viewed with critical suspicion.

Adequate data on the reliability of ratings are difficult to obtain. It is difficult to guarantee that ratings obtained from different raters at the same military establishment, factory, or office will be truly independent. At the worst the presumably independent raters may cooperate directly in preparing ratings. At the best it must be expected that both raters will be affected to some degree by the general reputation attached to each man in a particular military establishment or a particular job situation. The reputation a man has in a particular organization may be only in part a reflection of the actual quality of his performance and in part a reflection of the accidental circumstances which have occurred within the framework of that situation. Thus, a student pilot who happens to make a bad impression on the check pilot with whom he rides on his first check flight, possibly because of some incident entirely outside his own control, may be branded as incompetent and may be evaluated with biased eyes in subsequent flights because of that initial reputation. Only intimate knowledge of the situation prevailing in a particular post, office, or factory will indicate how serious the contamination of separate ratings may be. Evidence on this problem has been presented in Chapter 4.

The sources of unreliability in ratings of behavior during the performance of a specific task are threefold. First, we have the problem of subjectivity of observation. This subjectivity has been discussed in some detail above. As far as the rating a man receives is a function of the rater rather than the person rated, unreliability is introduced. Second, unreliability in ratings arises from the difficulty of setting a standard task to be observed and rated. Whether the performance is that of a co-pilot landing a transport plane, a salesman dealing with a prospect, or a teacher working with a class, the situation is likely to vary substantially



from one individual to the next. This source of irrelevant variance reduces the reliability of the resulting score. Finally there is the inconsistency of behavior of the individual from one limited sample of behavior to another. This of course is a general source of unreliability no matter by what means or instruments we make the observation.

## SUMMARY EVALUATIONS

In many cases, either because they are impractical or because they do not give a sufficiently broad base for judging the performance of the individual, observations of specific test periods are not satisfactory as criterion records. Obtaining these observations requires the development of special measuring techniques and the commitment of a substantial amount of personnel time to administer the test instruments to the employees being evaluated. In addition the observations are necessarily based on a somewhat limited sample of behavior and on a situation which may be somewhat artificial in relation to work upon the job itself. For these reasons, it often is necessary or desirable to fall back on evaluations of work that has been carried out over a period of time in the routine course of training or of performance on the job. Evaluations of this type, which involve a considerable period not under particular experimental control, are here called *summary evaluations*. They represent a summary score or judgment based on performance during a whole period of training or of work in the job itself.

In contrast with the specific evaluations of a limited behavior unit, which we have considered in the previous section, summary evaluations depend very largely on the routine records obtained in administering an organization. They grow out of the day-to-day contacts between employee and supervisor, the routine ratings used as a basis for evaluation and promotion, or the individual production records which serve as the basis for payment or accounting. Though it may be possible to set up special experimental conditions from time to time for personnel evaluation, summary evaluations must be based upon the routine conditions of observation and recording. The scope of the behavior included in them gives them a certain advantage over any more



standardized observation of a limited segment of performance; however, it must be recognized that they are subject to variation arising out of the whole gamut of conditions affecting records of performance on a job or the evaluation of performance.

Summary evaluations vary widely in their detailed characteristics, depending on the degree to which they are based on specific evaluations, the types of specific evaluations on which they are based, and the manner in which the specific evaluations are compounded. With regard to the degree of dependence on specific evaluations, at one extreme the summary may include nothing not already recorded as a specific evaluation of a defined segment of behavior. Average circular error in bombardier training was such a summary evaluation; it represented a simple averaging of bombing errors on a specified series of training missions. Most production records in industry are of this type, in that the total is a simple sum of production during specified shorter periods. At the other extreme, the summary may make no direct reference to any previous specific evaluations of behavior. Any rating procedure in which the rating is made without advance notice and without opportunity for further evaluation of the person rated is almost necessarily of this sort. One suspects that rating procedures in general, unless special provisions are made to the contrary, involve almost no reference to previous systematic observation or evaluation of the worker by the rater.

Summary evaluations may be based on various combinations of printed tests, objective performance scores, subjectively scored job samples, and rated job samples. These types of specific evaluations were discussed in the previous section. The qualities of the final summary evaluation will stem in part from the qualities of its component elements. Production records represent in a sense a compounding of specific objective performance scores. Most academic grades represent a composite of objectively scored written tests, which in that context may be considered objective performance scores, subjectively scored tests and exercises, and ratings of performances of different types in class and in connection with class exercises. Most employee ratings, in so far as they refer to *any* previous specific evaluations, probably

have reference to implicit ratings of particular performances by the employee.

The manner of combining specific evaluations into the final summary evaluation may represent a direct arithmetical compounding of the component scores, or it may represent a synthetic clinical judgment based on them in unspecified ways and giving an unspecified weight to each element. Direct arithmetical compounding is illustrated in such examples as those of bombardier circular error or industrial production record. Academic grades are also often based on such an objective procedure for combining partial evaluations. The outstanding example of the clinical synthesis of partial evaluations is, of course, the summary rating. Here, almost always, the final resultant bears no specifiable relationship to any specific previous evaluation of the individuals concerned.

In general, the conditions for a satisfactory summary evaluation are that it should be in large measure based on previous specific evaluations, the specific evaluations themselves should have desirable attributes as outlined in the previous sections of this chapter, and the procedures for combining the specific evaluations should be objective and well defined.

A summary evaluation of job performance can hardly be of much value unless it is based on previous observation of performance in specific situations. General after-the-fact impressions are notoriously untrustworthy and biased by irrelevant factors of general appearance, manner, and personal likableness. An illustration of this was provided by certain ratings obtained in the AAF for airplane commanders in operational training. The trainees were rated by their instructors on approximately ten traits. These same instructors indicated what they considered to be the importance of each of the traits for over-all effectiveness in the job assignment. Though "likableness" was consistently placed at the very bottom of the list in importance, it nevertheless fell at the top in terms of its correlation with an over-all rating. Though the raters disclaimed its importance, it still provided the chief basis for their over-all evaluation. Every supervisor makes certain observations concerning the work of people working for him. Unless some provisions are made to the contrary, however, these observations tend to be casual,

unsystematic, and to a considerable extent forgotten prior to any time for recording a summary evaluation. If the summary evaluation is to be based in any vital fashion upon actual performance on the job, it is important to develop specific procedures for maintaining records of performance from day to day.

It seems obvious that, the more relevant, accurate, and unbiased the specific observations have been, the more relevant, accurate, and unbiased will be the summary extracted from them. Finally, the values of the component evaluations can be maintained only if they are objectively combined. The possibility which clinical evaluation provides of allowing for external conditions is a small recompense for introducing into the final summary evaluation the unreliability and personal bias of subjective interpretation.

Summary evaluations, though differing in detail, can be discussed under four general categories. These categories are summary performance records, summary academic grades, summary ratings, and administrative actions. We shall now give some consideration to each of these categories.

### *Summary performance records*

In a number of job situations there are summary records of performance on one or another aspect of job duties. In the AAF these records were exemplified by average circular error records for the bombardier, percentage of hits in fixed gunnery for fighter pilots, and air-to-air target firing or gun camera scores for flexible gunners. In industry these performance records may be represented by weekly output for piece workers, total annual sales for life insurance salesmen, and other similar records of production.

These records will appeal strongly to the investigator in terms of their objectivity and apparent relevance to the ultimate criterion of the job in question. However, they may present in an even more acute form the problems of control of external conditions which were discussed in connection with specific performance records. The problem of controlling extraneous sources of variance, and consequently of attaining some minimum standard of reliability, is exaggerated in the present case by the fact that the data must be obtained under ordinary operating conditions rather than under the conditions of a special experiment. This

reduces the control of extraneous factors from that which *can* be obtained for purposes of research to that which typically is obtained in the normal course of training or operations.

In the AAF the reduction in experimental controls meant that all the factors of weather, equipment, other personnel, character of the target, etc., varied widely. In the industrial situation, the specific factors making for unreliability would be somewhat different, but their effect would be much the same. Differences in machines, in location, in supply of component parts, in severity of inspection, in day-to-day health and motivation would effect the record of the production worker. Variations in length of experience, in territory, in non-selling responsibilities, and the like would distort the records of the salesman. The question is whether under these circumstances it is possible to demonstrate *any* variation in performance consistently and unequivocally associated with the particular individual being studied. It may be reiterated that reliability need not be high to permit valuable research making use of a criterion, but it must be present.

In the AAF certain summary performance records appeared reasonably satisfactory from the point of view of reliability. For example, in one study of fixed gunnery scores of fighter pilots, an estimated reliability of 0.63 was obtained for 1200 rounds of air-to-air firing. This value is based on more than 1000 cases. On approximately the same group the reliability coefficient for 400 rounds of air-to-ground gunnery was estimated to be 0.59. In other situations, the reliability of performance records appeared much less satisfactory. A conspicuous instance of this was the average circular error in bombardier training in bombardier schools. A number of estimates of between-missions reliability were available for different bombardier classes. With these classes, using various analyses and various groupings of scores, coefficients were obtained, ranging from -0.08 to 0.37. The median of all the separate values was found to be 0.08. Thus, what appeared initially to be a decidedly relevant and quite promising criterion for evaluation of bombardier selection procedures turned out to be so very unreliable that its use as a criterion measure was almost impossible.

It is necessary also to consider systematic bias in the summary performance records for an individual. Whenever the different

factors of which we have been speaking vary from day to day and week to week, they produce the type of unreliable record that was reported above for bombardier training. However, when the factors remain uniform for a single individual throughout the period studied, they produce not unreliability but a biased score. The performance of the individual is consistent, but it is a function not merely of his own abilities but also of the specific situation in which he is placed. This tends to be more true of the civilian situation, in which a worker works consistently from day to day on the same machine or a salesman functions from month to month in the same territory. It may be possible to discover, measure, and allow for these biasing effects. However, the identification and allowance for them may be incomplete and inaccurate. When systematic biasing factors are present, the criterion may show satisfactory statistical reliability, but the reliability may arise from extraneous factors which have no relevance for the ultimate criterion.

### *Summary academic grades*

To a limited extent in industrial work, and to a very much greater extent in personnel problems connected with educational selection, academic grades merit consideration as possible criteria. In so far as a certain type of education is a prerequisite for a given job, such as that of the doctor or lawyer, success in the course of training has a type of relevance to the job in question. However, it is in connection with the quality of relevance that academic grades are most seriously suspect. Though a certain amount of unreliability and of bias in grades must be admitted, evidence from a variety of sources indicates that appreciable reliability remains in the final composite score. This reliability varies widely from situation to situation, depending on how extensive and how objective the procedures for evaluation in the training courses have been. However, it can be anticipated that in most cases the reliability of academic grades will be quite sufficient for personnel research purposes. The only problem in this regard is that of obtaining an adequate estimate of the size of the reliability coefficient. As indicated earlier a critical evaluation of the degree to which a criterion is being predicted depends on knowledge of the degree to which it is possible to



predict that criterion. The relevance of academic grades to ultimate success in any job is a much more serious question. Even in such professions as those of law, medicine, and engineering, it must be recognized that performance during training, and grades as an index of that performance, are only partial cues to eventual success in that job.

### *Summary ratings*

The criterion to which a personnel psychologist often turns, whether by choice or by necessity, is a rating of the employee by his supervisor. Rating systems are widely used in the armed forces, in civil service, and in industry as routine methods for evaluating personnel, and such ratings are often the most available criterion record. In many cases, no other type of record is readily available or conveniently procurable. Whether the employee is a file clerk or a vice-president of the company, it is often true that no production records or other objective evidences of performance have been maintained for him. Such records may have been left out because of practical considerations of the labor required to maintain them, or they may have been left out as entirely inappropriate to the job in question. In such a case, the psychologist must rely on ratings for an evaluation of the day-to-day performance of the employee on the job.

If ratings are to provide a relevant criterion measure of the individual two conditions must be met. The rater must be *willing* to rate the individual fairly, and he must be *able* to do so. A good deal of attention has been given to factors influencing the ability of a rater to rate his subordinate in accordance with the standards and intention of the agency collecting the ratings. Some of these points have been considered earlier in this chapter and in Chapter 3. However, not enough attention has been given to his willingness to do so.

It must be recognized that ratings are often required by a rather remote and impersonal controlling agency. Army Regulations, the Civil Service Commission, or top management require that periodic ratings be made at specified times and following a specified form. It is well known that these ratings influence promotion and preferment of individuals within the framework of the total organization. Now, one of the first maxims of

supervisor-subordinate relations is that a good supervisor stands up for the rights and interests of his subordinates. The loyalty of superior to subordinates may equal or surpass his loyalty to the distant and impersonal agency which has directed that ratings be made. Furthermore, he may feel himself in a sharply competitive situation with other subgroups of the organization. The ratings he gives his men may, therefore, reflect his eagerness to keep them contented and to win promotions for them on the one hand and his concern that his ratings come at least up to those of his competitors on the other. When this is true, instructions and exhortations from the agency using the ratings fall upon deaf ears, and the ratings crowd the upper end of the scale. This is well illustrated by the Army's wartime officer efficiency ratings. In this scale "very good" became a mark of disapproval, and "excellent" represented no more than the typical score.

The conflict of interest between the external agency requiring ratings and the small subgroup whose careers will be affected by them is almost inevitable in any organization in which real use is made of the ratings to determine the status and future of employees. When the rater is only to a limited extent concerned with rendering an impartial judgment in accordance with the standards and procedures set by the rating agency, his ratings cannot be expected to provide a satisfactorily relevant criterion measure. Almost any routine rating program may suffer from this defect.

General experience with the use of ratings, whether in the Armed Forces, civil service, or industry, confirms the personnel psychologist in the feeling that they are at best an unhappy choice. The choice may be, and often is, a necessary one, but it should be made with keen realization of the unsatisfactory solution which it presents. The limitations of rating procedures as applied to rating a specific segment of behavior have been described in a previous section. All these limitations are present in summary ratings and others as well.

It is an unfortunate characteristic of summary ratings that they are frequently not based in any clear way upon previous evaluations of specific behavior. The limitation may involve either the amount of specific information, the technique for synthesizing it, or both. In the extreme case, which is only too close to reality

in some instances, a summary rating represents an over-all judgment of an individual, rendered after an indeterminate period of acquaintance, given with absolutely no basis of previous systematic observation and evaluation of the individual. There are often no data in the form of observations, tests, or performance records. The rating represents merely an unguided, subjective, intuitive impression of the rater. In this case, the rating will obviously reflect personal bias and individual standards of judgment. If no other data are available, the rating may be expected to reflect nothing else but personal bias.

In this type of criterion, since biases are likely to be individual and are almost certainly unrelated to the ultimate criterion, the rating is also likely to have little to recommend it on the score of reliability or of relevance. An appearance of reliability may arise due to the general reputation factor which was discussed in connection with specific ratings. It may be suspected, however, that this will not hold up except within a limited group.

### *Administrative actions*

There are always a number of types of administrative actions which are taken with regard to personnel in any training or job situation. These often provide a summary evaluation of proficiency and merit consideration as possible criterion records. Logically, these are closely akin to the ratings just discussed. They represent in each case a judgment about an individual. However, in terms of their practical importance and of consequent possible differences in the manner in which they are prepared, these evaluations merit separate consideration.

The administrative decision which served most often as a research criterion for personnel psychologists in the AAF was the decision to graduate (or to eliminate) a man from a particular phase of training. Elimination because of lack of proficiency or for reason of fear or at the man's own request provided a readily available criterion measure which appeared to have some relevance both from the positive and the negative points of view. On the one hand, the skills and techniques which had to be learned in training provided the foundation for operations in combat. It seemed reasonable to believe that those who were particularly adept in learning the basic knowledges and skills

would, in general, be those who would be most proficient in later stages of operations. That is, training performance appeared to have some relevance to the ultimate criterion of combat performance.

At the same time, it was important to select for training those individuals who would complete and be graduated from that training and thus be available for assignment to combat duty. A man who was eliminated from training was by definition of zero value in that job specialty in combat. It may of course be argued that those men who were eliminated in training but who *could* have become successful in combat should never have been eliminated. It can be argued that training procedures and training eliminations were at fault and should have been changed. In the long run this is true. But working within practical limitations of time and an existing training situation, it may still be important to pick men who will succeed in that training situation. That is, performance during training appears to have some direct relevance for its own sake.

Other administrative actions were studied and used as criteria in the AAF to a lesser extent. These included both negative and positive actions. They are exemplified by re-evaluation and removal from flying by Flying Evaluation Boards, promotions, decorations, assignment to lead crews or other types of special duty, and removal from combat operations because of operational fatigue.

Practically any situation in education or industry will provide a series of administrative actions more or less analogous to the above. Men are dropped from college because of low academic records and other men are elected to honor societies. Men are released from their jobs because of incompetence or ineptitude. Other men receive promotions in pay or in rank. In specific instances there may be other actions which can reasonably be thought of as differentiating the more successful from the less successful on the job.

Practically all administrative actions imply a rating. These ratings differ from many others, however, in their immediate practical importance. Something is clearly being done on the basis of the rating. A man may be eliminated from training or from his job on the one hand, or promoted or put in a position

of critical importance on the other. On the basis of this, we may anticipate that the evaluation will be made more thoughtfully and conscientiously than when rating is merely an administrative chore. Relevant records will be consulted, testimony will be assembled and weighed, and the worst qualities of rating procedures will be somewhat mitigated. In some instances, such as election to an academic honor society, the basis for administrative action may be so rigorously specified that it takes on the degree of objectivity of an academic grade. It must be recognized, however, that most administrative actions do fundamentally imply ratings and that the limitations of rating procedures inhere in them.

### CONCLUDING STATEMENT

This chapter has explored the field of criterion measures. An attempt was made first to point out the crucial role of the criterion in any program of research for personnel selection and classification. The qualities needed in a criterion measure were indicated and their relative importance discussed. Then a number of possible types of criterion measures were presented for consideration, and attention was directed to the advantages and limitations of each. The general picture which emerges from this discussion is a somewhat sobering one. Really satisfactory criterion measures are not easily come by. Special testing procedures directed at getting measures of proficiency in a job require a very large investment of time and professional skill, and the results are often of limited scope and disappointing reliability. Routine records of performance must be scrutinized for the presence of external biasing factors. Obtaining ratings which have an adequate degree of relevance, reliability, and comparability from rater to rater represents a very difficult problem. The limitations which these difficulties represent must be recognized by any person engaged in selection research. His best efforts directed at the problem of obtaining satisfactory criterion data will be none too good.



## *The Estimation of Test Validity: Statistics of Validity*

In Chapter 5 we considered the problems involved in obtaining a criterion measure against which to validate selection measures. We shall now assume that a criterion score of some sort has been selected, even though that criterion may be an imperfect one, and inquire into the procedures for expressing the accuracy with which a particular test predicts that criterion. We shall first consider the general computational procedures to be used. Then we shall turn to certain special problems which arise, particularly with reference to restriction of range of ability in the group with which we have to deal.

### COMPUTATION OF VALIDITY INDICES

Indices of the validity of a test or other selection procedure may serve either of two purposes. On the one hand some indices provide a simple and graphic representation of the effectiveness of a procedure in discriminating between individuals of different degrees of success on the job in question. On the other hand some indices serve as analytical tools for relating the validity of the test to other statistics about it and to statistics about other tests. The goal of this analytical approach is to select most judiciously from among the available tests and to combine the chosen tests with most effective weights for a composite prediction of the criterion. It is probably always a mistake to expect any one type of statistical analysis to serve both purposes. In this chapter we are concerned only with indices which serve the second of the two purposes. Procedures for obtaining a quick pictorial representation of the effectiveness of a testing program will be considered in Chapter 11.

As tools for the analytical study of the validity of single tests and of the best manner of selecting from among them and weighting a group of them, the various types of indices of correlation appear almost uniquely valuable. The indices discussed in the following paragraphs all represent variations of the common product-moment correlation. When both the test score and the criterion measure are expressed as a score with a continuous distribution, the product-moment correlation can be computed. In some cases, however, either the predicting score or the criterion measure will be expressed as two or more discrete categories. Male-female is an example of a variable which might be used as a predictor and which is expressed as two discrete categories. Graduation versus elimination from training is a categorical criterion. In many cases these categories are twofold. In other cases, categorical data may be reduced to a twofold division, i.e., to the categories *X* and non-*X* in which non-*X* incorporates all other categories. The twofold or dichotomous variable therefore becomes of particular interest to us. We shall need to consider procedures which are appropriate when the test score is continuous and when it is dichotomous, and when the criterion is continuous and when it is dichotomous. We shall devote the following sections to considering the several combinations of criterion and of test score with regard to this factor of continuity or discreteness.

### *Continuous test scores—continuous criterion measures*

When both test and criterion scores are spread out over a continuum so as to provide a frequency distribution of score values, it is possible to express relationship between the two in terms of the Pearsonian product-moment coefficient of correlation. This situation prevails when we have the typical test of performance in reading, arithmetical skills, or mechanical comprehension as a testing instrument, and some score such as average academic grade, production rate over a period of time, or amount of commissions earned as a criterion measure. This index of relationship can, of course, show only the *linear* relationship between test score and criterion score, and if a curvilinear relationship seems to hold, more elaborate techniques must be applied. Some consideration will be given to non-linear

relationships later in the chapter. Formulas for the computation of the product-moment correlation will be found in any standard introductory text in statistics.

### *Continuous test scores—dichotomous criterion*

There are a good many cases in which criterion measures yield only a division of the criterion groups into two distinct categories. An instance of this type in the AAF was the dichotomy between graduation from training and elimination during the course of training. Graduation versus non-graduation would provide a dichotomy for almost any course of training, whether military, industrial, or academic. Promotion or non-promotion might be another dichotomy which could be used as a criterion measure. Similarly, discharge or release from employment as opposed to retention would represent such a dichotomy. In some cases, more than two discrete categories may exist, such as "promoted," "retained," and "discharged." It is always possible to reduce these cases to dichotomies by combining certain of the sub-groups, though some information is admittedly lost by such a procedure.

Two distinct indices of relationship are available when one measure is a continuous variable and the other is a dichotomy. These are based on two alternate assumptions as to the basic character of the variable which is available to us as a dichotomy. The first index is the *biserial correlation coefficient*. The biserial correlation coefficient is defined by the expression

$$r_{bis} = \frac{M_s - M_u}{S_t} \cdot \frac{pq}{z} \quad (1)$$

where  $M_s$  = mean score on the continuous variable of the "successful" or higher group on the dichotomy.

$M_u$  = mean score on the continuous variable of the "unsuccessful" or lower group on the dichotomy.

$S_t$  = the standard deviation on the continuous variable for the total group.

$p$  = proportion falling in the "successful" group on the dichotomized variable.

$q = 1 - p$ .

$z$  = the ordinate of the normal curve corresponding to  $p$ .

The derivation of the formula for the biserial correlation coefficient assumes that the variable represented in our data as a dichotomy is basically continuous and normally distributed. The break into two groups is considered to be arbitrarily imposed by some administrative condition and not to represent any fundamental or necessary break of the group at that point. Thus, some quality of "ability to fly an airplane" is assumed to underlie the administrative dichotomy of "graduation elimination." This ability is assumed to have a continuous and normal distribution, and the pass-fail dichotomy is considered to be representative of it. The proportion falling in the passing group is considered to be determined by the point on the continuum at which administrative policy at a particular time and place happens to have set the dividing line.

Implicit in the use of the biserial correlation coefficient is the feeling that what we are really trying to predict is amount or degree of the underlying quantitative variable. That is, the basic interest is not merely to predict whether an individual will fall above or below the particular dividing line which cuts the group into two fractions but also how far above or below he will fall. When the dividing line is thus arbitrary and a normally distributed variate (or one approximately so) does underlie the twofold division, and when we are interested in the accuracy of our prediction of that underlying variate, then the biserial coefficient of correlation is the appropriate index to use with a dichotomous criterion.<sup>1</sup>

<sup>1</sup> The computation of biserial correlation coefficients may be facilitated by using a variation of the formula presented above together with facilitating tables reported in J. W. Dunlap, "Note on Computation of Biserial Correlations in Item Evaluation," *Psychometrika*, 1, 51-58 (June 1936).

The biserial correlation is one specific case of a more general type of serial correlation, in which the criterion is expressed in two categories. In other cases the criterion may be expressed in three categories, such as "promoted," "retained," and "discharged," or in four or more categories. If the basic assumptions are tenable that the underlying variable is continuous and normally distributed and that the categories represent successive segments of that distribution, a generalized form of the serial correlation may be used. Formulas for the various cases of the serial correlations are given in N. Jaspens, "Serial Correlation," *Psychometrika*, 11, 23-30 (1946).

The alternate index to the above is the *point biserial correlation coefficient*. This is defined by the expression

$$r_{pbis} = \frac{M_s - M_u}{S_t} \sqrt{pq} \quad (2)$$

in which the symbols have the same meaning as for the biserial correlation coefficient. This index is derived on the assumption that the two categories are fundamentally and categorically distinct, and that all members of a single category are equivalent and are not to be distinguished with regard to the aspect under consideration. This is equivalent to saying that all individuals in the successful group are assigned a score of 1 on the criterion variable and all members of the unsuccessful group a score of 0. Thus, in the case of pass-fail in training it would be assumed either that the passing group and the failing group represented two categorically distinct groups (analogous to "male" and "female") or that the only concern of the research worker was to discriminate passers from failers in terms of the existing dividing line between the two groups. That is, if there were no concern for predicting degrees of excellence but only for predicting the single fact of pass or fail, the standard for passing being absolute and fixed, the point biserial coefficient of correlation would be the one to use.

Examination of the formulas for the two indices reveals that they differ only by the factor  $\sqrt{pq}/z$ . This factor equals about 1.25 when  $p = 0.50$ , 1.4 when  $p = 0.80$ , and 1.6 when  $p = 0.90$ . The biserial correlation is greater than the point biserial by this factor. When results from a number of tests administered to the same group of subjects are being compared, this factor necessarily is a constant. The same subjects are being used, so the number of "passers" and "failers" on the criterion remains the same. Under these circumstances, either index leads to exactly the same choice of the best tests and to the same assignment of relative weights to the two tests. It makes no practical difference which index is used.

When an experimenter proposes to compare or combine data from two or more groups in which the "successful" proportion  $p$  differs, then the choice of point biserial as opposed to biserial



correlation coefficients will generally influence the relative size of the numerical indices obtained for different tests. Since the biserial coefficient is strictly equivalent to a product-moment correlation coefficient, when the conditions of continuity and normality are satisfied, its value is independent of the value of  $p$ . The point biserial, however, is a function not only of the underlying relationship, but also of the evenness of the split into upper and lower groups. The more uneven the split, the lower the point biserial values tend to be. If the range of values of  $p$  in different groups is quite large, the relationship of biserial to point biserial may be seriously distorted. Thus, in one pilot training class in which 62 per cent were graduated, a biserial of 0.50 for the relationship of test composite to the graduation-elimination criterion would have corresponded to a point biserial of 0.39. In another class in which 88 per cent were graduated, the biserial of 0.50 would have corresponded to a point biserial of 0.31. That is, the same basic relationship of underlying variables would have given values differing by about 25 per cent if the point biserial had been used in these two cases. The difference is clearly quite an appreciable one.

The fact that a given basic relationship will yield the same value for the biserial correlation coefficient, no matter where the break is made dividing those passing and those failing, is one general reason for preferring this index of relationship. However, in each case one must examine the situation to see whether the assumptions underlying the biserial correlation coefficient are justified. These assumptions are, as has been indicated, (1) that the quality underlying the dichotomy in question is really a continuous variable and (2) that the underlying trait or quality possesses a normal distribution in the group.

The decision that the dichotomized variable really represents a continuum will be quite readily arrived at in many cases. Thus, in flying training it seemed reasonably clear that there was no sharp qualitative cleavage in underlying skill between those who were graduated from flying training and those who were eliminated. The difference was one of degree. This was substantiated by the wide variations in elimination rate from time to time and place to place. A pointed speech by a general

was enough to change the level at which the separation between the sheep and the goats was made. There was nothing natural or inevitable about the dividing line.

Justifying the assumption of a normal distribution for the trait underlying the dichotomy usually presents more of a problem. This is particularly true because the group available for experimental study is often pre-selected on some basis. That is, some testing or other selective procedure has been used to screen those admitted to the particular type of training or employment, and criterion data are available only for that fraction which passed the preliminary screening. It is possible and perhaps reasonable to consider that the skill in question would have been normally distributed *either* in the total group of applicants for the particular type of training or job *or* in the fraction selected for training on the basis of some type of screening procedure. However, if the screening had any validity at all (i.e., any relationship to later success), the distribution of skill could hardly have been normal in both cases. The more reasonable assumption usually is that the distribution was normal in the unselected population. This introduces a serious problem in using biserial correlations with screened or pre-selected groups. This problem will be developed later in the chapter, when the general problems of dealing with curtailed groups are considered.

### *Dichotomous prediction variable—continuous criterion*

Sometimes a variable used for prediction may be dichotomous. Thus, one might be using as a predictor the dichotomy "married" versus "not married" or the variable "high-school graduate" versus "not a high-school graduate." Here again, the biserial correlation coefficient and the point biserial correlation coefficient are two possible indices which may be computed. In deciding which one is appropriate to compute, we must determine which will give values that can appropriately be combined with the product-moment correlation coefficients for the remaining continuous variables. That is, when marital status is being considered as one predictor, we will usually be considering at the same time a number of other predictors, such as tests and ratings, each of which yields a continuous distribution of scores.

We must determine what weight shall be given to the dichotomous variable in relation to these other continuous variables.

The choice of index used will be dictated in this case by whether the score which we will eventually *use* for actual prediction by our dichotomous variable is still a dichotomy or is now a continuous score. For "married" versus "not married," we must use the point biserial. There is no possibility of obtaining a continuous score to substitute for that dichotomy. All we can do is throw the individual into one of two categories. For that case, the point biserial will be appropriate. Suppose, on the other hand, we have in our data the dichotomy "high-school graduate" versus "not a high-school graduate." If in our future use of this variable we can get information on the actual number of years of schooling so that we can supply a score on a continuum to replace the present twofold distinction, then the appropriate coefficient to indicate the predictive effectiveness of this variable relative to other continuous variables will be the biserial correlation coefficient. However, if all we shall know in the future is whether the applicant did or did not graduate from high school, the point biserial should be used in determining the weight to be given to this datum.

The argument of the previous paragraphs holds equally when the continuous variable with which the dichotomous predictor is being correlated is another prediction test. That is, biserial or point biserial will be used as in the case of a continuous criterion, depending on whether or not the dichotomous predictor will remain dichotomous in its final use for actual prediction purposes.

#### *Dichotomous prediction variable—dichotomous criterion*

Finally, we find some instances in which the variable being studied as a predictor and the indicator of success being used as a criterion are both dichotomous. This would be the situation if we were trying to predict graduation versus elimination in training from some item of information about an individual, such as his marital status. We must inquire concerning the available indices of relationship and the situations in which each is to be used.

We must recognize two interpretations of the criterion dichotomy, and two interpretations of the predictor dichotomy. The

criterion dichotomy may be thought of as an arbitrary division of a continuous normally distributed variate (case A). This is the case in which the biserial correlation expresses the relationship of a continuous test score to such a criterion. Or the criterion dichotomy may be thought of as a necessary and categorical division (case B). Here the point biserial expresses the relation of a continuous test score to the criterion. The test score dichotomy may be thought of as an incidental one which will be replaced by a continuous score before the test is used (case 1). This is the case in which a biserial correlation would be used in predicting a continuous criterion. It may be thought of as a necessary one which will have to be retained in any subsequent use of that test variable (case 2). Here the point biserial is used to predict a continuous criterion variable. There are four possible combinations of these interpretations, and a particular index of relationship appropriate to each.

In case A-1, in which both criterion and test are thought of as continuous, the *tetrachoric correlation coefficient* gives an index which corresponds to the product-moment correlation between the underlying continuous and normally distributed variables. It provides an estimate, computed from the dichotomies, of the correlation that would be obtained if the complete bivariate frequency distribution were available. Values for the tetrachoric correlation coefficient may efficiently be determined from a set of computing diagrams prepared under the direction of L. L. Thurstone.<sup>2</sup>

In case B-1 and case A-2, in which one of the two variables is thought of as basically continuous, or eventually to be used as a continuous variable, and the other as basically categorical, a coefficient analogous to the biserial correlation will be computed, treating the categorical variable as a point distribution. That is, the two values of the categorical dichotomy will be assigned the values 0 and 1, and the other dichotomy will be treated as a continuum. When values of 0 and 1 are assigned to the two categories, the formula for the biserial correlation coefficient

<sup>2</sup> L. Chesire, M. Saffir, and L. L. Thurstone, *Computing Diagrams for the Tetrachoric Correlation Coefficient*, University of Chicago Bookstore, Chicago, 1933.

takes a special form, which we shall designate biserial  $\phi$  ( $\phi_{\text{bis}}$ ). We get

$$\phi_{\text{bis}} = \frac{ad - bc}{z\sqrt{p'q'}} \quad (3)$$

where  $a$ ,  $b$ ,  $c$ , and  $d$  are the proportions in the four cells of the fourfold table;  $p'$  and  $q'$  are the percentages in the two categories of the real dichotomy (i.e., the one with the point distribution);  $z$  is the ordinate of the normal curve corresponding to the proportion  $p$  of individuals falling in the upper group on the artificial dichotomy.

In case B-2, in which both dichotomies are considered as categorical, the values 0 and 1 may be assigned to the two values of each of the dichotomies. If the product-moment formula is applied to these pairs of scores, where each score can take either the value 1 or the value 0, an index of relationship is obtained which has been designated the  $\phi$  coefficient. This coefficient is most simply expressed by the formula

$$\phi = \frac{ad - bc}{\sqrt{pq p' q'}} \quad (4)$$

where  $a$ ,  $b$ ,  $c$ , and  $d$  are the proportions in the four cells of the fourfold table and  $p$ ,  $q$ ,  $p'$ , and  $q'$  are the percentages in each category for each of the two dichotomies.

The cases in which each of these three formulas should be used have already been indicated in the preceding paragraphs. It should be noted that the four cases which have been identified (A-1, A-2, B-1, B-2) apply equally whether we are dealing with a predictor and a criterion or with two predictors.

### TAKING ACCOUNT OF RESTRICTION OF RANGE

In any personnel selection enterprise we are dealing with a certain population of applicants for the job or training in question. This population has a certain mean score and a certain variability of scores on any test applied. The statistics for the population of applicants are probably not the same as those for



the general public. However, the statistics for the population of applicants define the population with which we must deal, since they are the individuals from whom our selection must actually be made. The validity statistics in which we are basically interested, therefore, are the statistics which apply to that total population of applicants. In comparing two or more selection devices, we should compare them in terms of their validity for this general population of applicants.

Unfortunately, validity statistics often do not become available for a representative sample of this general population of applicants. We tend to *select* those who will be accepted for employment or training. As soon as some selection procedure operates to pick certain types of individuals from among those applying, the group for whom criterion records subsequently become available ceases to be representative of the general group of applicants. Their means and standard deviations on test scores are affected, and also the validity coefficients obtained for those test scores. In proportion as a high level of selectivity is in effect and in proportion as the bases of selection are closely related to the test scores being studied, the validity coefficients will be increasingly distorted. There is an essential incompatibility between using a selection procedure in practice and doing research upon it.

Unfortunately, this type of selection affects not only the absolute size of validity coefficients but also their relative size, so that the test which is really most valid as applied to the general run of applicants may appear to be one of the less valid in a group resulting from high standards of pre-selection. The reduction in the validity of a test within a selected group becomes greater the more closely the test correlates with the basis of selection.

The influence of selection upon the resulting validity coefficients becomes a very substantial matter where a high standard of selectivity exists. Data can be cited to illustrate this quite dramatically. In one project in the AAF Aviation Psychology Program, a special experimental group was studied to which all the usual selection tests were given but whose members

were admitted to pilot training no matter how poorly they did on the tests. Subsequently, biserial correlation coefficients against pass-fail in training were computed for (1) the total group who entered training and (2) those of the group who would have qualified for training in terms of the rather severe selection standards in effect toward the close of the war. For the composite aptitude score (stanine) and for certain illustrative tests validity coefficients were obtained as follows:

	TOTAL GROUP (N = 1036)	QUALIFIED GROUP (N = 136)
Pilot Stanine (Composite Score)	.64	.18
Mechanical Principles Test	.44	.03
General Information Test	.46	.20
Complex Coordination Test	.40	-.03
Instrument Comprehension Test	.45	.27
Arithmetic Reasoning Test	.27	.18
Finger Dexterity Test	.18	.00

It can be seen that where the restriction of the selected group is severe, amounting here to the exclusion of about 87 per cent of the cases, the changes in the resulting correlations are very striking indeed. The rather small size of the "qualified" group makes the results somewhat unstable. However, this degree of curtailment reduced the validity of the Pilot Stanine, which was the basis on which the "qualified" group was defined, by over 45 points. The first four tests, which entered rather heavily into the composite score designated Pilot Stanine and which therefore correlated quite highly with it, lost on the average 32 points. The last two tests, which had no weight in the Pilot Stanine for this group, lost on the average about 14 points in validity in the restricted group. The relative size of the validity coefficients in the restricted group gives little basis for judging the validity of the tests as applied to the complete population of applicants.

Though the degree of selectivity illustrated in this example is not likely to be equaled in many personnel research enterprises, the effects will be present to a degree whenever it is desired to carry on a research program at the same time that selection procedures are actually being used. If any intelligent use is to

be made of validity statistics from a restricted group, some statistical correction procedures are necessary to estimate what validity coefficients would have been obtained if it had been possible to obtain test and criterion data from a representative sample of all those to whom the selection devices were applied.

The general problem of inferring statistical parameters in a population from those which have been obtained in a sample when the sample has been selected or curtailed in some way with respect to the range of one or more variables has long been recognized. Karl Pearson<sup>3</sup> has treated the same essential problem in quite a different context and provided a solution which is applicable when the variables under study have a normal distribution in the general uncurtailed population. These formulas are about all that is available to us, and, since the values which they provide will almost certainly represent an improvement over the values found within the restricted group, it is appropriate that we consider them in some detail.

We may recognize three types of situations involving curtailment. These are differentiated in terms of the variable that has served as the instrument for curtailment and in terms of the variable whose standard deviation in the unselected population is known. The three situations will be characterized in turn, and the appropriate correction formula indicated.

*Case 1.* In this case, we are concerned with the correlation between variables 1 and 2. The basis for restriction of the curtailed group is variable 1. The values of the standard deviation are known for both the total (unrestricted) and the restricted group for variable 2. This situation would hold under the following circumstances. A research test, variable 2, had been given to a random sample of applicants and the standard deviation determined for that random sample. At some later date variable 2 had been correlated with the selection test (or the composite score used as a basis for selection), variable 1, in a group who had "passed" the selection test. The standard deviation of the research test (variable 2) would be known in both the

<sup>3</sup> Karl Pearson, "Mathematical Contributions to the Theory of Evolution—XI. On the Influence of Natural Selection on the Variability and Correlation of Organs," *Trans. Royal Soc. (London)*, Series A., 200, pp. 1-66 (1903).

general and the restricted groups. The correction formula then becomes

$$R_{12} = \sqrt{1 - \frac{s_2^2}{S_2^2} (1 - r_{12}^2)} \quad (5)$$

where  $R_{12}$  is the correlation between variables 1 and 2 in the unrestricted group.

$r_{12}$  is the correlation between variables 1 and 2 in the group that has been subject to selection or curtailment.

$S_2$  is the standard deviation of variable 2 in the unrestricted group.

$s_2$  is the standard deviation of variable 2 in the group that has been subject to selection or curtailment.

This particular case is not likely to be encountered often in practice.

*Case 2.* In this case, we are again concerned with the correlation between variables 1 and 2. Again, the basis upon which the curtailed group has been restricted is variable 1. In this case, however, we know the standard deviation of variable 1 in both the general and the restricted group. This situation arises under the following circumstances. We have administered some test or group of tests and obtained a score (variable 1) for each applicant. Then we have used the score as a basis for accepting only part of the applicants. For the accepted applicants we have correlated the scores on the test with some other score, such as a criterion measure (variable 2). We wish to know what correlation we would have obtained for the total group. The formula for correction becomes

$$R_{12} = \frac{r_{12} \frac{S_1}{s_1}}{\sqrt{1 - r_{12}^2 + r_{12}^2 \frac{S_1^2}{s_1^2}}} \quad (6)$$

where the meanings of the symbols are analogous to those of formula 5.

This case has a fair amount of practical significance because it is encountered whenever we wish to obtain an estimate of the validity of a selection procedure which we have actually been using as applied to the general group of applicants for the job category. Thus, in the AAF this formula was applied in correcting the validity coefficients for the stanines (composite scores) used in screening men for each of the air-crew specialties.

*Case 3.* In this case we are still concerned with the correlation between variables 1 and 2. However, we are now dealing with a case in which curtailment has been neither upon variable 1 nor variable 2 but upon some third variable. Data must also be available for the standard deviation of this variable 3 in both the general and the restricted group. This situation arises under the following circumstances. We have administered some test or group of tests and obtained a score on it (variable 3) for each applicant. On the basis of this score, we have accepted only a selected fraction of the applicants for the job specialty. For the accepted applicants, we have the correlation between two other variables (variables 1 and 2). These variables may represent a test and a criterion measure, two tests being considered as possible predictors, or two criterion measures. Most typically this correlation will be between a selection test (either one that forms part of the selection composite, or a new research test) and a criterion measure. We desire to estimate this correlation for the general unrestricted group.

In this case, the formula for correction becomes

$$R_{12} = \frac{r_{12} + r_{13}r_{23} \left( \frac{S_3^2}{s_3^2} - 1 \right)}{\sqrt{\left[ 1 + r_{13}^2 \left( \frac{S_3^2}{s_3^2} - 1 \right) \right] \left[ 1 + r_{23}^2 \left( \frac{S_3^2}{s_3^2} - 1 \right) \right]}} \quad (7)$$

The meanings of the several symbols are analogous with those in formula 5. The resemblance of the pattern of this formula to the standard formula for partial correlation should be noted. In this case, we are not seeing what will result if a third variable is held constant, but what will result if it is allowed to vary more widely. Thus the term  $\left( \frac{S_3^2}{s_3^2} - 1 \right)$  appears in this formula



in place of minus one as a multiplier of the correlational product terms.

There are certain instances in which one of the correlations used in the above formula may be available for the total rather than the restricted group. Thus, a research test may have been given to a general unselected group, and its correlation with the selection score may be based on this group. In this instance, the formula may be transposed into those terms. We get

$$R_{12} = \frac{r_{12} \sqrt{1 + R_{13}^2 \left( \frac{s_3^2}{S_3^2} - 1 \right)} + R_{13} r_{23} \left( \frac{S_3}{s_3} - \frac{s_3}{S_3} \right)}{\sqrt{1 + r_{23}^2 \left( \frac{S_3^2}{s_3^2} - 1 \right)}} \quad (8)$$

Case 3, in which restriction is indirect, imposed by indirect selection on the basis of some variable other than the ones being compared, appears by far the most common and most important one for any personnel selection research program. It is the situation that prevails whenever any score is being studied other than the particular one being used to screen the group. In any new research test, this is the situation which we will encounter. It is also the situation which we encounter when we start to make a critical analysis of the tests underlying any composite score.

So far, we have limited our consideration to the situation in which restriction of the group has been carried out on a single score. It has been assumed that magnitude of score on this variable is the only consideration determining whether or not an applicant will be accepted. No other considerations enter into the decision in any way. In practice, this idealized situation often (perhaps generally) does not prevail. The actual situation may deviate from the simple pattern which we have outlined in either or both of two ways. First, selection may have been based on specific reference to two or more scores. Second, selection may have been based on reference in unspecified and unspecifiable ways to various intangible and subjective factors which were never combined into a score. Of these two complications, the former represents a considerable increase in and complication of the statistical and computational labors of cor-

recting the obtained values; <sup>4</sup> the second represents an insuperable obstacle to any analytical treatment. When selection is based, as it often is, on a clinical judgment which combines in an unspecified and inconstant fashion various types of data about the applicant, and when this judgment is not expressed in any type of quantitative score, one is at a loss as to how to estimate the extent to which the validity coefficient for any test procedure has been affected by that screening.

<sup>4</sup> *Technical Note.* In the article by Pearson, cited in footnote 3, a derivation is given of the general formulas which are appropriate when the group has been curtailed on two or more variables. The same essential formulas for the general solution, but based on somewhat different assumptions and expressed in more convenient form, have recently been seen by the author in an unpublished manuscript by E. Reeve. The same formulas were independently arrived at by A. P. Horst. The basic assumptions of this derivation are:

1. The regressions of the non-restricted variables on the restricted may be considered rectilinear throughout the total population.

2. The variability of any given non-restricted variable is the same for each value of the restricted variables.

The formulas are presented in matrix notation, and the following symbols are used:

$x$  = a variable which is not directly restricted.

$a$  = a variable which is directly restricted.

$r$  = a matrix of correlations in the restricted group.

$R$  = a matrix of correlations in the unrestricted population.

$H$  = a diagonal matrix giving ratios of standard deviation ( $\Sigma/\sigma$ ) of the unrestricted to the restricted group.

$b$  = a matrix of partial regression weights (beta weights) in the restricted group.

Using this symbolism, it can be shown in matrix notation that

$$R_{xa} = R_{aa}H_a b_{ax}H_x^{-1}$$

where

$$R_{xx} = H_x^{-1}(r_{xx} - b'_{xa}r_{ax} + b'_{xa}H_a R_{aa}H_a b_{ax})H_x^{-1}$$

$$H_x = 1 - b'_{xa}r_{ax} + b'_{xa}H_a R_{aa}H_a b_{ax}$$

The matrix notation used presents extended series of operations quite compactly and has the additional advantage of suggesting work-sheet layouts and the order of operations for carrying out the necessary calculating procedures. It can be seen that the computations will be quite laborious at best when several variables are directly restricted.

## CURTAILMENT WITH A DICHOTOMOUS CRITERION

The formulas just discussed were all derived for product-moment correlations based on continuous variables. They are not strictly applicable to a biserial correlation obtained for a dichotomous criterion variable. This is because two incompatible sets of assumptions must be satisfied at the same time. On the one hand, if the formula for biserial correlation is legitimately to be used in the restricted group, the distribution of the trait underlying that dichotomy must be normally distributed in the restricted group. On the other hand, Pearson's formulas assume that the variable is normally distributed in the total population. The distribution can be normal in both cases only in the trivial case in which there is no correlation in the total population between the variable that has been the basis for restriction and the other test and criterion variables.

For the simple case of direct curtailment on a single variable (referred to as case 2 in the preceding section), a technique has been developed by Gillman and Goode,<sup>5</sup> which escapes the above dilemma. This is essentially a procedure for obtaining a least-squares estimate of the slope of the regression line from that part of the distribution which remains after curtailment. The procedure is as follows:

Let  $G$  = correlation estimated from this procedure (subsequently referred to as a  $G$  coefficient).

$f$  = number of individuals with score in the interval  
 $a \leq x \leq b$ .

$p$  = fraction of the above individuals falling in the upper  
 criterion group on variable  $y$ .

$u$  = abscissa value of the normal curve corresponding to

$$p = p_u.$$

$$X = \frac{z_a - z_b}{p_a - p_b}.$$

<sup>5</sup> L. Gillman and H. H. Goode, "An Estimate of the Correlation Coefficient of a Bivariate Normal Population when  $X$  Is Truncated and  $Y$  Is Dichotomized," *Harvard Educ. Rev.*, 16, 52-55 (1946).

Then compute

$$N = \Sigma f, \Sigma fX, \Sigma fX^2, \Sigma fu, \Sigma fXu$$

From these

$$A' = N\Sigma fXu - (\Sigma fX)(\Sigma fu) \quad (9)$$

$$D = \Sigma fX^2 - (\Sigma fX)^2 \quad (10)$$

Then

$$\tan \theta = -\frac{A'}{D} \quad (11)$$

$$G = \sin \theta \quad (12)$$

This computing procedure provides a technique for estimating the correlation in the total population of applicants for a job in the simplest case, in which the correlation is between a single variable, which has served as the basis for selection, and a dichotomous criterion. In this procedure, the assumption of normality of the dichotomous variable in the unrestricted population of applicants is still involved. However, no assumption must be made as to the nature of the distribution of this variable in the restricted group for which criterion data have become available. In this regard, the procedure that has just been described is to be preferred to procedures that require the computation of a biserial correlation coefficient in the curtailed group and the subsequent application of Pearson's correction formula.

No procedures analogous to the one just described are known for the case of indirect curtailment, which has been referred to as case 3. Unfortunately, this case is likely to occur most frequently and to be of most importance in any program of personnel selection research. Whenever a new research test is being studied, or whenever an analysis is being made of tests which contribute to a composite score used for selection, the curtailment is of this indirect character. Of course, these situations are crucial to any critical analysis of and improvement of a test battery. In these cases, there appears to be no completely satisfactory technique for estimating the validity of the predictors of a dichotomous criterion.

As we indicated, the Pearson formulas for correcting for curtailment are not strictly applicable to biserial correlation coefficients. No analytical solution is available to indicate the direction or amount of error when these formulas are applied to

biserial coefficients. In the AAF Aviation Psychology Program, one study was made of synthetic artificial data to obtain empirical evidence on the extent and direction of the errors involved. This was carried out only for the simplest case (case 2 above), in which direct restriction upon a single variable is involved. Tables of synthetic data were prepared assuming normal distributions for each variable. The test scores were expressed in single digit form, ranging from 1 to 9, in which each score value corresponded to a range of one-half standard deviation. Tables were prepared assuming a correlation of 0.50 in the population and for various elimination rates. From these tables, curtailed groups were set up by eliminating first the 1's, then the 1's and 2's, etc. Biserial correlations were computed from the curtailed groups and corrected values were determined by applying formula 6. From this analysis the following findings emerged:

1. In the cases studied, which were so designed that the dichotomy in the curtailed group was more uneven than in the total group, the correction formula uniformly tended to yield an underestimate of the correct value for the total group.

2. The amount of underestimation increased as the amount of curtailment increased and as the unevenness of the dichotomy in the unrestricted population increased. In the most extreme case studied, in which the restricted group consisted of only the upper 40 per cent of the total group and in which the population split of graduates and eliminees was in the proportion 90:10, the formula underestimated the true value by some 20 per cent.

The above results refer only to the less crucial case 2 of correction for direct curtailment. No analogous investigations were made of the extent of error in applying the Pearson formula to biserial correlations in the more significant case 3, in which the curtailment takes place on some third variable. Some error is undoubtedly involved in this case, but for lack of any better procedure it is probably desirable that formula 7 be used. If any substantial amount of selection has operated, the values obtained after applying this formula will almost certainly be more accurate than the raw values based upon the curtailed group. The discussion in the preceding pages has indicated that the problems involved in inferring population values from data



available from a more restricted sample of accepted applicants are far from being completely solved. Some of the problems which remain can be briefly indicated.

1. No formula is available which is strictly applicable to biserial correlations in the case of indirect curtailment on a third variable or in the case of curtailment on more than a single variable.

2. The formulas all assume that it is possible to specify and get measures of the variables which have served as the bases of curtailment. No procedure is available for dealing with the effects of selection where that selection has been based in unspecified degree on a number of unmeasured subjective factors which have not been combined into a score and used in an objective manner.

### NON-LINEAR RELATIONSHIP

The correlational procedures discussed so far in this chapter have all assumed that the relationships between the variables being studied are linear. It has been assumed that score on the criterion variable will show a progressive and uniform change for each unit of change in a test score or other predictor. There may be situations in which this does not happen. We can conceive of situations in which a certain amount of some particular ability is an asset to the man engaged in a particular job, but where further increments in that particular ability may not be helpful. Thus, a carpenter may need a certain minimum of numerical ability in order to carry out simple calculations involved in his work, but he may not benefit further from a high level of numerical ability. An elementary-school teacher may need at least average intelligence to comprehend the situations with which she must deal, but she may not benefit further from abstract intelligence of a high order. Many other cases will come to mind for which it seems at least plausible that test scores are related to the degree of job success through only part of the range of scores.

The initial attack on the prediction of any criterion of job success is normally in terms of the standard linear measures of

relationship which have been discussed. However, it is always desirable for the research worker to examine the hypothesis of linear relationship, to see whether it does in fact apply to the data at hand. When both test and criterion measure are continuous variables, so that the product-moment correlation coefficient is used to obtain a measure of relationship between them, the linearity of the relationship may be tested quite simply. This is done by computing the *correlation ratio* and comparing the correlation ratio with the correlation coefficient.

The correlation ratio for predicting  $y$  from  $x$ ,  $\eta_{y|x}$ , is defined as the ratio of the standard deviation of column means to the standard deviation of the total distribution of  $y$  scores. It can be obtained by the formula

$$\eta_{y|x} = \sqrt{\frac{\sum n_k \bar{Y}_k^2 - N \bar{Y}^2}{N \sigma_y^2}} \quad (13)$$

where  $\bar{Y}$  is the mean of the total distribution of  $y$  scores.

$\bar{Y}_k$  is the mean of the  $k$ th column of the bivariate frequency distribution.

$n_k$  is the number of cases in the  $k$ th column.

$N$  is the total number of cases.

$\sigma_y$  is the standard deviation of the distribution of  $y$  scores.

When the relationship between the two measures is exactly linear, the correlation ratio and the correlation coefficient will exactly coincide. In all other cases, the correlation ratio will be greater than the correlation coefficient. This is due to the fact that the correlation ratio is raised by any variation from column to column in the bivariate distribution—not merely a regular or linear change. In evaluating the departure of data from linearity, some test of the discrepancy between  $r$  and  $\eta$  must be applied to determine whether the departure from linearity in the sample is more than could reasonably have arisen as a sampling fluctuation from a population in which the relationship was actually linear. Fisher<sup>6</sup> develops a method of testing the goodness of fit of a regression equation to observed data making use

<sup>6</sup> R. A. Fisher, "The Goodness of Fit of Regression Formulae," *J. Roy. Statistical Soc.*, 85, Part IV, 597-612 (1935).

of the chi-squared distribution. In the present instance, in which a linear regression is being tested, a value of  $\chi^2$  is obtained from the expression

$$\chi^2 = (N - k) \frac{\eta^2 - r^2}{1 - \eta^2}$$

In this expression,  $N$  is the number of cases in the sample and  $k$  is the number of columns in the bivariate frequency distribution. The table of  $\chi^2$  must be entered with  $k - 2$  degrees of freedom ( $n = k - 2$  or  $n' = k - 1$ ).

If the above test shows that it is necessary to reject the hypothesis of linear regression, it may then prove advantageous to determine a curvilinear relationship between test and criterion. For exploratory work, a smooth curve may be fitted by eye to the set of column means, and the fit of this curve to the set of points represented by the column means may be judged by inspection. A more elegant approach is to fit a second degree parabola or some other mathematical function to the set of points by the method of least squares and to test the goodness of fit of this curve to the set of points by the chi-squared test. The test is analogous to that described above for the linear relationship. The method of least squares as applied to curve fitting will not be discussed here. It can be found in standard statistical texts. In particular, Peters and van Voorhis<sup>7</sup> give attention to curve fitting as applied to just those problems of correlation and prediction which we are considering here.

If, for some one variable, it is found that a non-linear relationship gives a significantly better prediction of the criterion than a linear one, the scores on that variable must be transformed into new score values, using the equation which expresses the relationship between that variable and the criterion. Thus, if the equation for predicting  $y$  from  $x$  is of the form

$$\tilde{y} = a + bx + cx^2$$

<sup>7</sup> C. C. Peters and W. R. van Voorhis, *Statistical Procedures and Their Mathematical Bases*, Chapters 11 and 15, McGraw-Hill Book Co., New York, 1940.

we can substitute for  $x$  a variable  $u$  which is a function of  $x$  defined also by the expression

$$u = a + bx + cx^2$$

That is,  $u$  is the best estimate of  $y$  which can be obtained for this type of quadratic function in  $x$ . The values of the function  $u$  will have substantially a linear relationship to the criterion  $y$ , and these transformed  $u$  values can be analyzed in combination with other predictor variables to arrive at the best combined prediction of the criterion.

When the criterion is a dichotomous one, such as graduation-elimination, one additional complication is introduced into the study of linearity of the relationship of test to criterion. In this case, the relationship which we shall be able to study is the relationship between score on the predictor and percentage falling in the passing group on the criterion. We shall be concerned to inquire whether the change in percentage passing as the score increases corresponds to an essentially linear relationship.

The additional complication is introduced by the fact that the percentage scale is not itself a linear scale. Assuming a normal bivariate frequency surface, which is the assumption made if we use biserial correlation coefficients, any percentage of individuals passing can be translated into an abscissa value on the base line of the normal frequency curve. Relatively small percentage differences at the extreme percentage values represent rather substantial scale differences, whereas percentage differences near the 50 per cent value correspond to relatively small scale differences. In any study of linearity of relationship it is first necessary, therefore, to translate percentages into scale values in terms of the normal curve. These scale values can be plotted against test score and studied to determine their linearity or non-linearity. If they appear non-linear, some other type of mathematical function can be examined with view to getting a more adequate expression of the relationship. However, the tests of significance of departure from linearity reported for the product-moment correlation cannot be directly applied in the case of biserial correlations. It may be necessary to rely on inspection to decide whether the departure from a linear relationship is sufficiently

marked and sufficiently regular to suggest the desirability of expressing the relationship by some more complex type of function.

The possibility of encountering a significantly non-linear relationship should always be borne in mind. However, one should not count on finding such relationships. For example, in the study of many different tests for several different job criteria in the AAF, no convincing evidence was found for the existence of non-linear relationships.



## Combining Tests into a Battery

Most types of jobs call for a number of different aptitudes. Factors of intellect, skill, interest, and personal adjustment enter in a complex combination to determine how successful the worker will be. It is rarely possible to assess this complex pattern of traits with a single type of test material. Different types of testing techniques are required to assess the several traits making for success or failure. As soon as two or more separate test scores are available for use in selecting personnel for a particular job specialty, the personnel psychologist is faced with the problem of how to combine the information from the several scores so as to provide the most effective selection of personnel for the job.

There may be some ambiguity about the concept of *most effective selection*. In some cases, it may be important never to miss any of the best individuals. In other cases, as perhaps in the selection of automobile drivers, it may be important that every person selected should come up to a minimum standard of competence. As a general working definition, however, we may consider the most effective procedure to be that which yields a selection of personnel for the job who show the highest average performance on the job in question. If we have a criterion score of job performance, that selection procedure will be considered best which yields a group of the required number of employees for whom the average of the criterion scores is a maximum.

How shall we combine the scores from a number of tests in order to get the best prediction of job success? There are a great many possible ways of combining scores, and we shall consider some alternative possibilities later in the chapter, but the most generally useful one is to make a *linear combination* of the scores. This means that for each test some constant is determined. The constants are in general different for each test,

but the same constants are used for every individual. The composite score for each individual is the sum of his test scores, after each test score has been multiplied by its appropriate constant. That is,

$$\tilde{y} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (1)$$

where  $\tilde{y}$  = predicted job success, expressed in standard score units;  $x_1, x_2, \dots, x_k$  = score on tests 1, 2,  $\dots$ ,  $k$ , expressed in standard score units; and  $\beta_1, \beta_2, \dots, \beta_k$  = the multiplying constants for tests 1, 2,  $\dots$ ,  $k$ .

The problem is to determine the values for the constant multipliers,  $\beta_1, \beta_2, \dots, \beta_k$ , that will give us the most effective composite score. These weights, which are applied to the separate test scores to obtain the best prediction of the criterion, are called regression weights, and the equation based on them is the *multiple regression equation* for predicting the criterion score  $y$  from the several tests,  $x_1, x_2, \dots, x_k$ . That set of multipliers or weights for the separate tests yields the best score, in the sense discussed above, for which the discrepancies between composite score and measure of actual success on the job are the smallest.

Since there are always a great many ways of constructing composite scores so that the plus and minus discrepancies balance each other, that condition does not provide an adequate definition of a "best" composite score. We shall desire further that the absolute size of the discrepancies, or some function of them, shall be a minimum. The most satisfactory function to use for this purpose is  $(y - \tilde{y})^2$  and we shall specify that

$$\Sigma(y - \tilde{y})^2 = \text{a minimum}$$

This may also be written

$$\Sigma(y - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_k x_k)^2 = \text{a minimum} \quad (2)$$

Choosing weights so that the sum of the squared deviations of predicted from attained success is a minimum is known as the method of least squares. This method has widespread application in experimental science.

Evaluating a minimum requires some elementary calculus. We must take the derivative of the above expression with respect to each of the  $\beta$ 's. This gives us  $k$  equations of the type

$$\begin{aligned}
\Sigma 2x_1(y - \beta_1x_1 - \beta_2x_2 - \dots - \beta_kx_k) &= 0 \\
\Sigma 2x_2(y - \beta_1x_1 - \beta_2x_2 - \dots - \beta_kx_k) &= 0 \\
&\vdots \\
\Sigma 2x_k(y - \beta_1x_1 - \beta_2x_2 - \dots - \beta_kx_k) &= 0
\end{aligned}
\tag{3}$$

These may be expanded into the form

$$\begin{aligned}
2\Sigma x_1y - 2\beta_1\Sigma x_1^2 - 2\beta_2\Sigma x_1x_2 - \dots - 2\beta_k\Sigma x_1x_k &= 0 \\
&\vdots \\
2\Sigma x_ky - 2\beta_1\Sigma x_1x_k - 2\beta_2\Sigma x_2x_k - \dots - 2\beta_k\Sigma x_k^2 &= 0
\end{aligned}
\tag{4}$$

Since all scores are considered to be in standard score form, so that

$$\sigma_{x_1} = \sigma_{x_2} = \dots = \sigma_{x_k} = 1$$

these equations are equivalent to the equations

$$\begin{aligned}
\beta_1 + \beta_2r_{x_1x_2} + \dots + \beta_kr_{x_1x_k} &= r_{x_1y} \\
\beta_1r_{x_2x_1} + \beta_2 + \dots + \beta_kr_{x_2x_k} &= r_{x_2y} \\
&\vdots \\
\beta_1r_{x_kx_1} + \beta_2r_{x_kx_2} + \dots + \beta_k &= r_{x_ky}
\end{aligned}
\tag{5}$$

The above  $k$  equations, the "normal" equations, serve to define those values of the  $\beta$ 's which give the best linear combination of the separate test scores—best in the least squares sense, in that the sum of the squares of the differences between predicted success and actual measure of success will be a minimum.

In equations 5 we have a set of  $k$  linear equations in  $k$  variables. These equations must be solved to determine the regression weights  $\beta_1, \beta_2, \dots, \beta_k$ . The solution of a set of simultaneous equations is a straightforward process, but it becomes very laborious if the equations number more than three or four. For this reason special computing routines have been set up to systematize the process of computing and especially to provide frequent checks upon the accuracy of computing, so that errors may be discovered promptly and promptly corrected. In addition, procedures have been devised for arriving at the solution by a series of approximations. Two efficient routines for solving

the set of simultaneous equations are presented in Appendix A. The first is a condensed version of the Doolittle solution especially suitable for use with certain types of computing machines which permit the convenient summing of a series of multiplications. The second is a method of successive approximations which modifies in some details a procedure described by Kelley and Salisbury.<sup>1</sup> The iterative method does not provide a rigorous and completely objective solution, as the other methods do, but it places considerably less exacting demands upon the computer and often saves a good deal of time. The characteristics of the method are discussed further in Appendix A.

Whichever computing method is used, the solution of the normal equations 5 yields a set of values for  $\beta_1, \beta_2, \dots, \beta_k$ . These are the values for which the correlation of predicted success with actual success is a maximum and for which errors in prediction are a minimum. The correlation between the best weighted composite of test scores and the criterion variable is the *multiple correlation*. It is given by the expression

$$R = \sqrt{\sum_{i=1}^k \beta_i r_{x_i y}} \quad (6)$$

If some set of weights,  $A_1, A_2, \dots, A_k$ , other than the exact regression weights has been used to obtain a composite score, the correlation between the composite score and a criterion measure is given by the expression

$$R = \frac{\sum_{i=1}^k A_i r_{x_i y}}{\sqrt{\sum_{i=1}^k \sum_{j=1}^k A_i A_j r_{x_i x_j}}} \quad (7)$$

Formula 7, which is the general formula for the correlation between one measure and a weighted sum of other measures, is useful in estimating the correlation with a criterion which will result from any given set of test weights which may be suggested by rational considerations of some sort.

<sup>1</sup> T. L. Kelley and F. S. Salisbury, "An Iteration Method for Determining Multiple Correlation Constants," *Jour. Amer. Stat. Assn.*, 21, 282 ff., (1926).

The multiple correlation serves as an index of the degree to which a test battery is being successful in predicting a criterion. The square of the correlation gives the percentage of variance in criterion score which is predicted by the test battery. If we assume that the desired outcome of testing is a test battery which picks those individuals who will show the best average performance on the job, the multiple correlation is a direct measure of the success of the test battery.

The  $k$  equations given in equations 5 and the values for  $\beta_1, \beta_2, \dots, \beta_k$  which result from them are derived on the assumption that all scores are expressed as standard scores. These weights are, therefore, the weights to be applied to *standard scores* on each of the tests. If, as is usually the case, test results are obtained and used in raw score form, with different standard deviations for the different tests in the battery, then the weights must be adjusted to take account of these differences in standard deviation. Since

$$z_1 = \frac{x_1}{\sigma_1}, \quad z_2 = \frac{x_2}{\sigma_2}, \quad \dots, \quad z_k = \frac{x_k}{\sigma_k}, \quad z_y = \frac{\tilde{y}}{\sigma_y}$$

where  $z_1, z_2, \dots, z_k, z_y$  are scores in standard units and  $x_1, x_2, \dots, x_k, \tilde{y}$  are scores in raw score units, we have

$$\beta_1 \frac{x_1}{\sigma_1} + \beta_2 \frac{x_2}{\sigma_2} + \dots + \beta_k \frac{x_k}{\sigma_k} = \frac{\tilde{y}}{\sigma_y}$$

$$\beta_1 \frac{\sigma_y}{\sigma_1} x_1 + \beta_2 \frac{\sigma_y}{\sigma_2} x_2 + \dots + \beta_k \frac{\sigma_y}{\sigma_k} x_k = \tilde{y}$$

The weights to be applied to the several tests, when test results are expressed in raw score units, are therefore

$$\begin{aligned} b_1 &= \frac{\beta_1}{\sigma_1} \sigma_y \\ b_2 &= \frac{\beta_2}{\sigma_2} \sigma_y \\ &\vdots \\ b_k &= \frac{\beta_k}{\sigma_k} \sigma_y \end{aligned} \tag{8}$$



Since it is not the absolute numerical values of the weights which are important, but their relative size, we can multiply all the above weights by any convenient constant, such as  $\sigma_j$ . Thus, we can use as weights to be applied to raw scores

$$\begin{aligned} B_1 &= \frac{\beta_1}{\sigma_1} \\ B_2 &= \frac{\beta_2}{\sigma_2} \\ &\vdots \\ B_k &= \frac{\beta_k}{\sigma_k} \end{aligned} \tag{9}$$

or any other convenient set which is proportional to the above.

Let us review briefly the steps to be taken in selecting the set of weights which will give the best linear combination of test scores for predicting a given criterion. First, the intercorrelations of all the tests and the correlations of each test with the criterion measure must be determined. These must be appropriately corrected for any curtailment, following the procedures discussed in Chapter 6. From the table of intercorrelations and validities the values of  $\beta_1, \beta_2, \dots, \beta_k$  must be calculated, for which one of the methods presented in Appendix A may be used. For use with raw scores, the weights must be corrected to take account of the respective standard deviations, in accordance with formulas 8 or 9. These then constitute the best set of weights. How effective they are can be seen by the multiple correlation given by formula 6.

## FACTORS DETERMINING MULTIPLE CORRELATION

The effectiveness of the composite score which can be derived from two or more tests depends on two considerations. These are (1) the validities of the single tests and (2) the correlations between the tests. The fact that the validity achieved with a test battery is dependent on the validities of the tests composing the battery seems quite directly apparent. The role of the separate test validities becomes more apparent if formulas 5 and 6 for multiple regression and multiple correlation are examined critically. We can see then that the validity resulting from a

battery of tests is a direct function of the validities of the separate tests. If all the separate test validities were multiplied by some constant, at the same time leaving the correlations between the tests unchanged, the multiple correlation which could be obtained from the tests would be multiplied by the same constant. Thus, two tests with validities of 0.25 and 0.30, respectively, and an intercorrelation of 0.20 will give a multiple correlation of 0.36. If the validities are 0.50 and 0.60 and the intercorrelation remains 0.20, the multiple correlation becomes 0.72.

The practical importance of test intercorrelations is less immediately apparent, but no less real. The effects of varying intercorrelations between tests can be illustrated quite simply and dramatically by computing the multiple correlation for varying values of test intercorrelations. Let us assume that we have several tests, each of which has a correlation of 0.30 with a criterion. Let us next assume that all the intercorrelations of these tests are first 0.00, then 0.10, then 0.30, and finally 0.60. The values of the multiple correlation which can be obtained from various numbers of tests that conform to these specifications are shown in Table I. This table shows how severely

TABLE I. EFFECT OF INTERCORRELATION ON MULTIPLE CORRELATION

(Multiple correlation resulting from different number of tests, when validity of each test is 0.30 and intercorrelations are uniform and at several different levels)

No. of Tests	Size of Intercorrelations			
	.00	.10	.30	.60
1	.30	.30	.30	.30
2	.42	.40	.37	.34
4	.60	.53	.44	.36
9	.90	.67	.48	.37
20	*	.79	.52	.38

\* It is mathematically impossible for twenty tests all to correlate 0.30 with some measure and still have zero intercorrelations.

intercorrelation between tests limits the gain from the addition of further new tests. Where the intercorrelations are 0.00, the addition of three comparable tests to one whose validity is 0.30 permits a composite score with a validity of 0.60. Where the intercorrelations are 0.60 the possible increase is only from 0.30

to 0.36. The limiting effect of intercorrelation becomes greater as the number of tests is increased.

The contribution that any single test can make to the effectiveness of a battery for predicting some criterion is a function both of its correlations with the criterion and of its correlations with the other tests. There are two ways in which a test can add to the validity of an existing test or pool of tests. It may provide a measure of some additional valid variance, or it may serve as a suppression test. The test that measures new valid variance will show a substantial validity coefficient but will not have high correlations with existing tests. (Usually the validity coefficients and intercorrelations are both positive in sign, but occasionally a test may be encountered in which the score values are reversed, so that both validity and intercorrelations are negative.) An effective suppression test is one which measures only the non-valid variance which appears in an otherwise valid test. It will show a high correlation with the test for which it is a suppressor but a low correlation (though often a positive one) with the criterion. This type of test will enter into a composite prediction with a negative weight. Its function is to partial out the non-valid variance from an otherwise valid test, making the composite score a purer measure of the valid factor. An illustration of the operation of the suppression variable will make these points clearer.

Imagine a written test for mechanics made up of a number of questions about the use of machine tools and about shop procedures. Let us suppose that this test has a validity of 0.40 for some criterion of performance as a machinist. Suppose we have a reading test of comprehension of literary selections which has a validity for the machinist criterion of 0.10. Let us assume that the correlation between the two tests is 0.60. If we then determine those weights for combining the two tests which will yield a composite score with the highest possible correlation with the machinist criterion, we find that the weights are 0.53 for the machinist test and  $-0.22$  for the reading comprehension test. The multiple correlation for the two tests is 0.44. In this case, the reading test has served to partial out certain verbal and reading factors which were serving to attenuate the validity of the machinist test. The composite score becomes, then, a purer measure of the mechanical factor, or whatever we want to call

the factor or factors making for validity in the machinist test, and has higher validity than the machinist test alone, in which this factor was diluted by others. The suppression test has increased validity not by broadening but by narrowing the scope of our measurement.

Most of the tests in a battery usually find their way there because of their positive validity for the criterion. Suppression tests are the exception. Thus, in the AAF air-crew testing program there was only one occasion on which a test was used as a suppression variable and given a negative weight. The accumulation of further data showed this test to have a higher validity than preliminary results had indicated, and the negative weight was found not to be justified. Analysis of any limited sample of data is likely to show small negative weights for a few tests. However, these may very possibly represent sampling fluctuations for tests whose weight in the population would be zero. This problem of sampling fluctuations will be discussed presently. In practice, the personnel psychologist should probably hesitate to weight negatively a test which has positive validity for the criterion. The statistical evidence should be supported by rational grounds for expecting that test to serve as a suppression variable for other tests before a negative weight is introduced. In test invention and construction, however, the concept of the suppression variable presents interesting possibilities.

### ALTERNATE PROCEDURES FOR USING TEST DATA IN SELECTION OF PERSONNEL

So far we have discussed one procedure for using tests in personnel selection. This is the procedure of determining a weight for each test, and then giving each applicant a score which is the weighted sum of his separate test scores. The composite score for each individual is then a *linear combination* of the scores on the single tests. The procedure for arriving at the composite score is a completely objective arithmetical one, involving each test as a first-degree term in a simple weighted sum. Those individuals are selected for the job who have the highest composite scores, or those individuals are considered satisfactory for the job who have composite scores above a given minimum.

There are several other ways to use test scores and other data in selecting personnel. We shall describe three and consider their limitations and purported advantages. The procedures are the following:

1. Test scores may be combined according to some algebraic function, but this may not be a simple linear function of the single test scores. That is, the composite score may involve the square or cube of test scores, or the product of tests 1 and 2, or some more complex function. The tests are combined algebraically, but not as a simple sum. This will be designated the non-linear function method.

2. Each test score may be used entirely independently of the others to establish a minimum qualifying score. An individual will be considered qualified only if he comes up to the minimum on *each one* of the tests. He will be rejected if he falls below the minimum on *any one* of the tests. This will be designated the multiple-cutoff method.

3. The test results may be used in a non-mathematical subjective manner by an individual presumed to have a high level of insight into patterns of individual personality and into the needs of the job. This person makes a clinical judgment of each applicant, on the basis of all the available evidence. This judgment provides the basis for accepting or rejecting that applicant. This will be designated the clinical method.

### ***Non-linear function method***

This method is a generalization of the procedure for selecting personnel on the basis of a composite score derived from a linear combination of test scores. It is similar to the linear regression equation technique in that a single composite score is obtained which is an algebraic function of the scores on the separate tests. It differs in that the function is not restricted to a linear one but may involve any more complex function of the test scores. The non-linear function method reduces to the standard regression equation technique in the limiting case in which all but the first-degree terms in the prediction equation disappear.

Since it includes the linear case but is much more general, the non-linear function method has all the theoretical advantages, plus greater flexibility. Theoretically it is possible to take account of situations in which success is a function of the joint



presence of two traits rather than either of them separately, i.e., where success depends upon the product of two scores rather than the sum of them. Other more complex situations can also theoretically be taken care of. Thus, if success in a criterion score were being predicted from performance on two tests, it would be possible to work with a prediction equation which included the several possible quadratic terms. We would get an equation of the form

$$\bar{y} = a_1x_1 + a_2x_2 + a_3x_1^2 + a_4x_2^2 + a_5x_1x_2 \quad (10)$$

Using this equation as a starting point, it would be possible to solve for the  $a$ 's, using the method of least squares, just as we did in the linear case. However, we should now have five normal equations instead of two, and these would involve such terms as  $\Sigma x_1^4$  and  $\Sigma x_1^3x_2$ . The computational labor would obviously become extremely heavy, even for only two variables. If any substantial number of variables were involved, the introduction of anything other than linear functions of the test scores would obviously result in an impossible computational burden.

The use of a more general non-linear function of the test scores for predicting success is theoretically quite attractive. It makes more general and flexible the approach to personnel selection through use of an algebraic function of the several test scores. In a few rare cases, where tests are few, cases abundant, the project of great importance, and the situation such as to give some logical support to a non-linear function for prediction purposes, this type of analysis may be worth trying. However, despite its theoretical attractiveness, it is not generally useful for analysis of tests for personnel selection.

### *Multiple cutoff method*

In this method, the research worker undertakes to set a minimum qualifying score on each one of the tests, and an individual is accepted only if he qualifies on every one of the separate tests. The several qualifying scores are set with the objectives of (1) qualifying the proportion of applicants needed to meet current personnel demands and (2) qualifying those who have the greatest promise of success on the job. In this procedure the individual receives no composite score. His test record is merely

inspected to see whether he qualifies on all tests, and he is placed into one of two categories, the qualified or the disqualified.

In evaluating the use of multiple cutoffs, we may ask a number of questions. We must ask first whether the procedure provides as effective a selection of those likely to succeed on the job as does the regression equation. We must then inquire whether the statistical analyses are efficient and practical. We must finally ask whether the method provides a flexible and convenient procedure for personnel selection and classification. Let us examine these points in turn.

The manner in which the multiple cutoff procedure selects those to be accepted for a given job can be compared with that of the multiple regression equation most effectively in the case of two test variables. Let us assume that we have administered two tests to a group of subjects and that we wish to determine from the results on these two tests the most accurate prediction of success on some criterion, such as success in medical school. The joint distribution of scores on the two tests can be shown by a bivariate frequency distribution and might fall into some such pattern as shown in Figure 1.

In the multiple regression technique, a linear combination of the two test scores is determined and a single aptitude score is computed

$$\tilde{Y} = b_1X_1 + b_2X_2$$

If a particular standard of aptitude is required or a specified proportion of applicants is needed on the job, a minimum qualifying value of  $\tilde{Y}$  is established. This is represented in Figure 1 by the line  $a-a$ . All individuals falling below and to the left of line  $a-a$  are disqualified, and all those falling above and to the right are qualified. The slope of line  $a-a$  is a function of the relative weights,  $b_1$  and  $b_2$ , of the two tests in the combined aptitude score. The position of line  $a-a$  is a function of the standard set to qualify for training.

For the multiple cutoff technique, separate minimum scores are set for each of the two tests. There is no really analytical way of establishing the two critical scores below which an individual shall be disqualified. The procedure is necessarily a trial-and-error one. Different combinations of score values for  $X_1$  and  $X_2$  must be tried. For each of the selected combinations

the research worker must determine (1) the percentage of cases disqualified by using this combination of cutting scores and (2) the amount of difference in average criterion score for the accepted group and the rejected group. That combination of cutting scores will be chosen which (1) yields the proportion of accepted applicants which fits the supply of candidates on the

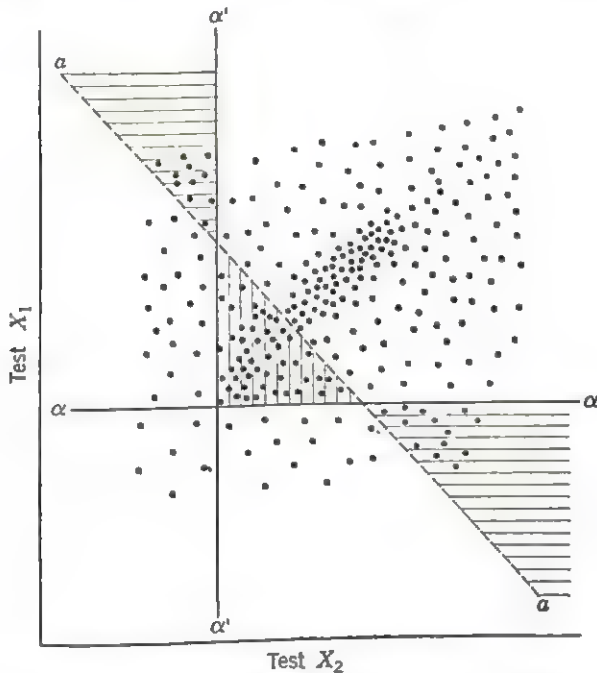


FIGURE 1. Comparison of multiple regression and multiple cutoff procedures for personnel selection.

one hand and the demand for job placements on the other and (2) makes the sharpest discrimination in criterion score between those who are accepted and those who are rejected. A statistical technique for testing this sharpness of discrimination has been developed, called the multiple chi test.<sup>2</sup>

The cutoff scores on the two tests are represented in Figure 1 by the lines  $\alpha\alpha$  and  $\alpha'\alpha'$ . The difference in actual results by the

<sup>2</sup> R. Franzen, *Method for Selecting Combinations of Tests and Determining Their Best "Cut-Off Points" to Yield a Dichotomy Most Like a Categorical Criterion*, CAA Division of Research, Report 12, Washington, D. C., March 1943.

two procedures is represented by the lined areas in the figure. Those individuals falling in the area with horizontal lines will be accepted by the multiple regression analysis but not by the multiple cutoff procedure. Those falling in the vertically ruled area will qualify by multiple cutoff but not by multiple regression. All other cases will be classified in the same way by both methods. Multiple regression will accept those who though below one cutoff are substantially above the other and will reject those who are just barely above both cutting scores.

The one case in which we would expect those who were selected by the multiple cutoff procedure to surpass in criterion performance those selected by multiple regression is that in which the relationship of one or more of the tests to the criterion is sharply non-linear. If there is some unique critical score on a particular test below which all or most applicants do poorly on the job and above which a much smaller proportion do poorly on the job no matter what their other qualifications, then a procedure which determines that point and establishes a fixed cutoff at that point undoubtedly has advantages. However, in so far as a continuous and approximately linear relationship exists between score on each of the tests and the criterion score of success on the job, no basis exists for choosing a uniquely desirable cutting score. In this case, we may expect that superiority in other tests will offset a small deficiency in one particular test, and any rigid requirement for a particular test seems not to be defensible. Even when non-linearity of relationship of a test to the criterion is found, some function of the original test scores can often be found which does have a linear relationship to the criterion. This problem is discussed in Chapter 6. If this function is substituted for the raw test scores, the multiple regression procedures often prove entirely satisfactory.

In discussing the problem of non-linearity, it should be noted that one may often expect some non-linearity to appear in empirical data from a limited sample of cases. The crucial issue is whether this same non-linearity appears in subsequent samples. Any critical empirical test of the results for multiple cutoff procedures must be based on the effectiveness of the cutoff values as applied to a new sample distinct from the one that was used to determine the cutting scores.

A second problem concerning the multiple cutoff technique is the feasibility of the statistical analyses which are required. As indicated, a trial-and-error procedure is necessary to determine the cutting scores to be used. Various possible cutting points must be tried for each of the tests, and these must be tried out in all their possible combinations. When only two or three tests have been used, the number of combinations is not very large, but this number multiplies very rapidly as the number of tests is increased. If it were proposed to examine the appropriateness of no more than three different cutting scores on each of ten different tests, the total number of combinations for which results would have to be analyzed would be some 59,000. Under these circumstances, the method obviously breaks down under its own weight.

A third problem concerns the adaptation of multiple cutoff procedures to varying conditions of supply and demand. One of the realities that any personnel program must face is that either the supply of job applicants or the number of job vacancies or both is likely to change from time to time. A selection procedure should be capable of adaptation to these changed conditions. Where a single composite score serves as the basis for qualifying applicants, flexibility is almost unlimited. By merely raising or lowering the minimum acceptable score, any desired proportion of applicants can be qualified. No other change need be made in the procedures. No such ready adaptation of the multiple cutoffs is possible. To determine the best combination of cutting scores for a new required percentage of acceptance, a rather complete re-examination of all the original analyses is required.

A fourth problem concerns the adaptation of multiple cutoff procedures to the problem of positive classification, as distinct from selection. The classification problem will be discussed more fully later in the chapter. Essentially, the problem is to decide to which one of several job specialties a person should be assigned. The problem is no longer one of accepting or rejecting the applicant but rather of making the most effective use of each individual in some one of the several jobs. Since the multiple cutoff procedure yields no quantitative score indicating *degree* of suitability for any job, it provides absolutely no basis for



determining for which one of several specialties an individual is best qualified. It is difficult to see how the technique could be adapted to this purpose. The impossibility of using the results for classification will represent a major limitation of the multiple cutoff method of analysis in many personnel research projects.

### *Clinical method*

The methods described so far are strictly objective, as applied to any individual, and the critic may contend that they are mechanical and lifeless. A number of measures are obtained for each individual, and the scores resulting from these measures are treated in a specified and uniform way to determine whether the applicant shall be accepted for the job. It should be noted that the measures may include observational records, ratings, and other data which are arrived at through the medium of the human observer. However, once these data have been obtained they are reduced to scores and treated as such.

It has often been contended by the clinically minded that the strictly objective methods previously described lose a good deal of valuable information which could be used by a skilled clinician in judging the suitability of each individual. The clinician, it is asserted, can make use of various types of data about the individual which emerge in a clinical interview but which are not readily reduced to scores for quantitative treatment. Furthermore, he is in a position to interpret the *unique pattern* of test scores for an individual and is not bound to the rigid and mechanical procedure of weighting which is represented in a regression equation.

We must distinguish here between the use of an interview and a rating resulting from it as a datum in an objective selection program and the use of data from tests as raw material for a synthetic final clinical judgment. The usefulness of interview and rating as a datum can be determined in the same way that the usefulness of any test score is determined. The validity and intercorrelations of the rating may be calculated, and it can be treated as one variable in a multiple regression equation. Its value will be determined by whether it does or does not get an appreciable regression weight. The present problem is to decide whether an essentially clinical treatment of test and other data represents a desirable approach to the selection problem.

The only advantage in a clinical evaluation of a set of scores is that it permits combination of the scores in other than a linear manner. It permits a maximum of flexibility in that any pattern, no matter how complex or unique, may be recognized and weighted. For this extreme flexibility to be an advantage, it is necessary (1) that special patterns and combinations of tests, not well represented by a linear combination of scores, be important for success on the job and (2) that there be clinicians available who have the insight to discover those special patterns and the skill to recognize them whenever they appear. We may well be somewhat skeptical on both counts, but especially on the second. It represents a severe demand on a clinician's insight to expect him to discover better ways of using test scores than will be given by the best linear combination of those scores, and then to be consistent in identifying and interpreting those patterns when they reappear.

In the final analysis, the effectiveness of clinical interpretation of test scores can be judged only by experiment. The clinician may be supplied the scores and instructed to make his best prediction of success on the job, based upon the scores and all other information available to him. The validity of his predictions may be compared with the validity obtained from the regression equation. Whatever the results, however, they show only what *one* clinician or group of clinicians was able to do in *one* personnel situation. Findings on this problem will necessarily be extremely specific. The dependence of the procedure upon the particular persons involved is one of the serious weaknesses of the clinical method. Even if it can be demonstrated that the clinical method of assessing data is valuable as applied experimentally by certain persons with specialized training and abilities, there is still no guarantee that those values can be maintained week in and week out in a continuing personnel selection program.

## LENGTH OF A TEST BATTERY

When validities and intercorrelations are available for a battery of tests, the question usually arises as to how many of the tests should be retained for final use in routine selection. It is possible to take one, two, three, or any number less than the

complete set of tests and find how good a prediction can be obtained from this shorter battery. The multiple correlation for the shorter battery is often almost as good as that for the complete battery. (It cannot be quite so good in the specific sample analyzed unless the test or tests which have been omitted actually receive zero weight in the regression equation. Any weight given to the omitted variable would be given only if it increased the multiple correlation.) The problem is to determine whether (1) the loss from dropping out some of the tests is a real loss, in the sense that it would be maintained in a new sample, and (2) the loss is enough to overbalance the increased efficiency and simplicity of a shorter battery.

A comparison of batteries of different numbers of tests is quite straightforward if the comparison is restricted to the sample of cases used in the current analysis. Either one may start with the complete battery and drop tests one at a time, dropping each time the test with the smallest regression weight, or one may start with a single test and build up the battery one test at a time, adding always the test that will make the greatest addition to the existing pool.

In either the Doolittle or the iterative methods of computing regression weights given in Appendix A, it is possible to drop successive variables from the solution. The iterative procedure lends itself particularly well to this process, because all that must be done is to make the weight for the specified variable zero, and then make adjustments by successive approximations to the weights of the other variables. The validity of the abbreviated battery will drop slowly at first as single variables are eliminated, and then more and more rapidly as variables with larger regression weights are eliminated.

A technique for the successive addition of variables has been developed by Wherry.<sup>3</sup> This procedure starts, of course, with the most valid single test. Then one determines which variable will add most to the validity of this one test. (In essence, one determines the partial correlation of each test with the criterion when the first test is held constant.) The weights for the two variables in combination and the multiple correlation for that combination of tests are computed, and one then determines

<sup>3</sup> R. J. Wherry, in W. H. Stead and C. P. Shartle, *Occupational Counseling Techniques*, American Book Co., New York, 1940, pp. 245 ff.

which test will add most to the original two. The process is continued step by step, adding at each step the test that will give the greatest increase in the multiple correlation based on the existing pool of tests.

The problem that arises in either of these approaches is that of when to stop dropping or adding tests. At what point does the gain from an additional test variable become so small that it represents no real gain but only the increase that should be expected by chance from the addition of another parameter to the prediction equation?

In any given sample, correlations will deviate from population values due to the vagaries of sampling. Even if the correlations with the criterion are all zero in the total population, some prediction of that criterion will be possible within the sample. The multiple correlation will be a function of the number of predictor variables. The more predictor variables, the higher the multiple correlation which one may expect to derive from them, in the absence of any genuine relationship. At what point does the improvement in prediction represent merely this fact?

If one is comparing a set of  $k$  tests with a set of  $k + 1$  tests, where the  $k$  tests are taken at random from the  $k + 1$ , a test of the comparative size of the multiple correlation resulting from the two sets is possible. We can interpret the correlation from the larger battery as really representing a gain in accuracy of prediction only when

$$\frac{1 - R_{k+1}^2}{1 - R_k^2} < 1 - \frac{1}{N - k - 1} \quad (11)$$

where  $R_{k+1}$  = the multiple correlation for  $k + 1$  variables.

$R_k$  = the multiple correlation for  $k$  variables.

$N$  = the number of cases in the sample.

Unless the above inequality holds, the increase in  $R_{k+1}$  over  $R_k$  must be thought of as representing only the effects of sampling operating in the larger number of variables.<sup>4</sup>

The above test applies only when the  $k$  variables are chosen at random from the  $k + 1$ . This, however, is not the situation

<sup>4</sup> The derivation of the formula, attributed to Churchill Eisenhart, may be found in Paul R. Rider, *An Introduction to Modern Statistical Methods*, John Wiley & Sons, New York, 1939, pp. 126-128.



when we successively remove variables from or add variables to a battery. In that case our selection is very definitely systematic. Starting with the complete battery, we remove first those variables with the smallest regression weights, or starting with a single variable we add first those that make the largest contribution to the initial test or tests. This sequential process fully capitalizes on chance fluctuations in validities and intercorrelations in the sample. The test that is most valid in a particular sample is so in part because of its genuine validity, but in part because *in that sample* that test happened to have a large positive deviation from its true population value. Thus, the tests that are added first tend to have positive deviations in the present sample, and those added last tend to have negative deviations. The apparent validity of the first few tests will be considerably inflated, and the apparent contribution of the last few sharply minimized. The above formula takes no account of this fact. Wherry's formula which undertakes to correct the multiple correlation for the shrinkage to be expected as one goes to a new sample is also inappropriate when variables are added in the systematic way which we have been discussing.<sup>5</sup>

If a large number of tests are given to a relatively small group of subjects and correlated with each other and with a criterion variable, the sampling fluctuations among the correlation coefficients will almost always be sufficient to produce some few test correlations that will give a substantial prediction of that criterion in that sample. However, in addition to any true relationship that may exist between the test variables and the criterion in the parent population, a prediction based on a few chosen tests capitalizes on the chance fluctuations in validities and intercorrelations of the variables in the specific sample. In general, the variables that are retained from a larger group will

<sup>5</sup> R. J. Wherry, in W. H. Stead and C. P. Shartle, *Occupational Counseling Techniques*, American Book Co., New York, 1940. The formula is

$$\bar{R}^2 = 1 - K^2 \left( \frac{N - 1}{N - M} \right)$$

where  $\bar{R}$  is the "shrunk" multiple correlation.

$K$  is the multiple coefficient of alienation.

$N$  is the number of cases.

$M$  is the number of tests.



be not only those that are most valid but also those that show the most favorable fluctuation from their true population value in the current sample. The smaller the number of variables selected to be weighted, the more premium is placed upon favorable chance fluctuations in those particular variables retained and weighted. It must be anticipated, therefore, that the multiple correlation obtained by weighting only a picked few of the tests in a battery will show a marked shrinkage, substantially *greater* than when the regression weights are based on all the tests. This is the exact reverse of the situation represented by Wherry's shrinkage formula. The author knows no accurate analytical formulation to express the amount of shrinkage to be expected from the weighting of a specified number  $k$  selected from a battery of  $n$  test variables, and no procedure can be offered for estimating this effect with precision.

In the absence of a solution of the basic statistical problems, the personnel psychologist is thrown back largely upon rule-of-thumb procedures and practical judgment in determining the number of tests to keep in his battery. One empirical approach is to determine intercorrelations and regression weights on one experimental population, and then apply these to an independent new sample as a test of their effectiveness. In this way, an unbiased test may be obtained of the effectiveness of weighting different numbers of test variables. However, practical considerations of the time and facilities available for testing are often the deciding factors in determining how extended a routine battery shall be.

### ADDITION OF TESTS TO AN EXISTING BATTERY

After a personnel selection or classification program has been established and has been in operation for some time, a nuclear battery of tests will have been developed for routine testing operations. If the program includes an active, continuing research aspect, new tests will be continuously in the process of development and validation. Whenever validation data are obtained for a new test, a decision must be made as to whether that test should be put into the routine testing battery, either as

an addition or as a replacement for some test already in the battery.

The final statistical basis for this decision must be a complete analysis of the regression weights and multiple correlation resulting when the new candidate is added to the battery or substituted for a test already included therein. However, this analysis is a laborious and time-consuming enterprise. Also it requires that all the correlations of the new test with the rest of the tests in the present battery be computed. There are certain types of preliminary analysis which may give some cues as to (1) which tests are the most promising ones to administer for validation purposes and (2) which of the validated tests show the most promise of making a real addition to the validity of the existing battery. These are based on analysis of the correlations between the new research test and either the tests already in the battery or the composite score derived from them.

The first type of analysis involves the determination of the *uniqueness* of a new research test. A test's uniqueness is defined as the systematic variance in the test which is distinct from and cannot be predicted by the other tests in the battery. The variance in a set of test scores can be divided into three parts: (1) common variance, (2) chance variance, and (3) specific or unique variance. The percentage of common variance in a test is defined as the square of the multiple correlation between that test and the rest of the tests in the battery. The percentage of non-chance variance (i.e., common plus unique) is equal to the reliability coefficient for the test. The percentage of unique variance is therefore given by the difference between the reliability coefficient of the new test and the square of its multiple correlation with the rest of the tests in the battery.

Information about the uniqueness of a test may be of value in the negative sense, in that a test with very little uniqueness cannot possibly add much of anything that is novel to the variance already covered by the battery. Therefore, it can hardly contribute to the total validity of the battery. It may possibly give a purer measure of some function already appearing in two or three other tests, and this greater purity may be of value for classification purposes, as will be pointed out later in the chapter

Unless a test has a fairly substantial percentage of unique variance, however, we can hardly expect it to add appreciably to the validity of the existing composite score or scores. Such a test would usually not be very promising to give for validation.

A second analysis which is possible as soon as a test has been administered to a preliminary sample of job applicants is based on the correlation of the test with the existing composite score. When we have determined the correlation of the new research test with the present composite score, and if we also know the validity of that composite score, we can immediately determine what validity would be required in the test to add a specified amount to the validity of the present composite score. The formula for this computation<sup>6</sup> is

$$r_{ck} = r_{ks}R_{c \cdot t} \pm \sqrt{a(a + 2R_{c \cdot t})(1 - r_{ks}^2)} \quad (12)$$

where  $a$  = the specified increase in the multiple correlation.

$r_{ck}$  = the validity required of test  $k$  to achieve the increase  $a$ .

$R_{c \cdot t}$  = the multiple correlation of the battery, excluding test  $k$ , with the criterion.

$r_{ks}$  = the correlation of  $k$  with the composite score when test  $k$  is excluded from the battery.

If some standard for increase in validity is set a priori, formula 12 can be used to determine how valid a test would have to be in order to yield that increase. Thus, if the present composite score had a validity of 0.50 and a new test correlated with that composite score to the extent of 0.60, then, in order for the test to add as much as 0.02 to the validity of the present score, it would have to have a validity above 0.41. (If the validity were below 0.19, the desired increment would also result. In this case, the new variable would be negatively weighted and serve as a suppression test.) Of course, this evidence as to the size of the required validity can serve as only a rough guide in establishing priorities for test validation. However, by combining this indication of the validity which a test would *have to have* with one's best professional judgment as to the validity

<sup>6</sup> This formula was developed by A. P. Horst while he was serving in the Army Air Forces. So far as the author knows, the derivation has not been published.

which it is *likely to have*, a better rational guide is provided for assigning priorities to tests for validation.

A third analysis is possible after validity data become available for a research test. Though still preliminary to analysis of the complete matrix of correlations of the new test together with the tests already in the battery, this is much more definitive than the pre-validation analyses which we have just been discussing. This analysis is based on test validity, composite score validity, and correlation of test and composite. It consists merely of determining the multiple correlation with the criterion of the test and the present composite. The increase of this multiple over the validity of the present composite shows how much increase in validity could be obtained from the new test if it were added to the present composite *without changing the internal weighting of the composite*. Assuming that the other tests in the composite are already correctly weighted, this provides a minimum estimate of the gain to be obtained from the new test. It is a minimum estimate because internal changes in the existing composite would be made only if they resulted in further increase in the multiple correlation. Of course, if the tests already in the battery are not optimally weighted, it is possible that the same increment in validity which is provided by the new test might be achieved completely or in part by re-weighting the tests already in the battery. Though the estimate resulting from this procedure is presumably a minimum estimate, we may anticipate that it will not be much of an underestimate. It seems unlikely that very much further gain will result from changing the weights of the other tests.

For those tests which appear promising by this last analysis, complete correlational analysis to determine their weights and contribution to the validity of the existing set of tests is in order. This requires the determination of the correlation of the new test with each of the tests in the current battery. It also requires the matrix of intercorrelations of the existing tests and, of course, the test validity coefficients. The iterative procedure for determining test weights, presented in detail in Appendix A, may be used quite advantageously for this analysis. The previous weights for tests already in the battery usually provide a good initial approximation from which to make further corrections, and

this procedure is likely to lead to rapid convergence of the weights upon their final values. The speed of the method makes it practical to analyze a number of research tests singly and in combination and to determine what they add to the present battery. This addition is seen in the increase in multiple correlation when one or more research tests are added to or replace tests already in the battery.

The problem of deciding when the data on a new test are sufficiently promising to merit adding the test to a current battery or substituting it for a test already in use is not a simple one. Data on a research test are usually much more limited than data on tests which have been used routinely for selection over a period of time. As a result, the estimates of validity for the research test are likely to be less securely based than those for battery tests, and to be subject to a greater amount of sampling error. Furthermore, these sampling errors are not random but are definitely biased. Those tests which gave the highest validity coefficients and the lowest correlations with the other tests will be considered as candidates for addition to the battery. These high validity coefficients represent in part genuinely high validity in the tests in question, but in part they represent a plus fluctuation from the true population value in the sample which was analyzed. In so far as a sampling fluctuation is involved, the validity may be expected to drop in a new sample. The extent to which size of the validity coefficient is attributable to sampling fluctuations is very difficult to estimate. It is a function not only of the size of the group on which the validity coefficient was computed, but also of the number of tests from which the particular research test was selected. The greater the number of tests, the more likely one is to get large chance deviations in one or two of them.

In view of these factors, no exact mathematical statement seems possible concerning the increment in validity to be required before a new research test is added to the battery. It seems reasonable to be rather conservative, especially if the validity of the current battery is fairly high and is determined for a substantial number of cases, and to add new tests and especially to replace existing tests only when the evidence in



favor of the research test is very convincing. It also seems very desirable that a new research test which has been added to a test battery be subject to continuing scrutiny, and that new validation data continue to be gathered for it.

### COMBINING DATA FROM PARTIAL CRITERIA

As was indicated in the discussion of criterion measures in Chapter 5, for almost any job specialty there are a number of types of data which represent possible criterion measures of success in the job. These criterion measures are of different types and become available at different times during the course of training for or service in the job. Each usually has some logical or statistical claim for relevance to the ultimate criterion of success in that job. Thus, in pilot training in the AAF one type of criterion was supplied by elimination from individual flight training, and this could be further subdivided into the stages of primary, basic, and advanced training. Criteria for fighter pilots in operational training included such data as air-to-air and air-to-ground gunnery scores, accidents, reclassification to non-flying status, and various types of ratings. Combat provided such additional criterion material as promotions, decorations, reported planes shot down, casualties, and reclassifications. These types of criterion information all appeared to be in some degree relevant to judging the success of a particular individual. Many of the correlations among these separate criterion measures, when it was possible to compute them, were quite small. Different criterion measures are likely not to show high correlations, partly due to unreliability but partly because they measure different aspects of the complex total which is job success and measure it together with different irrelevant variables. If validation data are available for two or more different partial criteria, we then face the problem of combining the validity statistics for the several partial criteria to arrive at weights to be used in selecting personnel for assignment to this particular job specialty.

There is one possible analytical, mathematical approach to the combination of data from partial criteria. The computation

of a canonical correlation<sup>7</sup> provides a determination of the maximum prediction of a weighted group of criterion measures from a weighted group of prediction variables. In this procedure, weights are assigned to criterion variables to yield a composite criterion score and to predictors to yield a composite predictor score. Both sets of weights are so chosen that the correlation between the two resulting scores is a maximum. The result is the mathematically unique maximum prediction of some composite of those criterion measures from some combination of those tests. However, though this solution is mathematically the best, we may question whether it is the best in any real or practical sense. The criterion measure that receives a heavy weight in the composite, because it can readily be predicted, is not necessarily important according to the best available professional judgment. In the allocation of weight to criterion measures judgment rather than statistics must be the final court of appeal.

The alternative to an analytical solution of our problem in terms of maximum prediction is to fall back on professional judgment for the determination of the weights for the partial criteria. The weights may be applied specifically to the criterion variables. Thus, a group specializing in the field may decide that amount of insurance sold should be given twice as much weight as rating by the district manager. This would mean, then, that the composite criterion score for each individual would be so constructed that amount of insurance sold would receive twice the effective weight given to district manager's rating. (The effective weight must take account of the standard deviations of the two distributions.)

Often the groups on which the different criterion measures are available are not the same. In this case, it is not possible to obtain composite criterion scores for single individuals. The weights then have to be applied to the validity coefficients obtained for the partial criteria or even to the regression weights obtained for them. It can be shown that the *relative* weights obtained for several prediction variables eventually receive are the same at whatever stage data for the partial criteria are combined. Since

<sup>7</sup> H. Hotelling, "Relations between Two Sets of Variates," *Biometrika*, 28, Parts 3 and 4, pp. 322-377 (1936).

the relative weights are the significant consideration, the point at which the combining is done is immaterial.

The important thing in combining data from partial criteria resolves finally into obtaining the soundest possible judgment as a basis for assigning weights to the several criterion variables. The judgment should probably represent both the specialist in the job and the specialist in measurement. The specialist who knows the job well should be best able to pick out the behaviors that are important evidences of success in the job specialty. The measurement specialist is best able to evaluate the technical characteristics of the scores which can be gathered as evidence of each of those behaviors. Their combined judgment should provide an evaluation both of the importance of the behavior and of the adequacy of the measure of it. These two considerations jointly can determine the weight given to each criterion variable.

### ROLE OF NON-STATISTICAL FACTORS IN WEIGHTING TESTS

The discussion so far in this chapter has been concerned almost entirely with statistical procedures for arriving at the weights to be assigned to test variables when the tests are being used to select personnel for a job. However, the immediately preceding section has suggested that judgment, as well as statistics, enters into the final weights. At this point it is appropriate to consider the role of reasoning, judgment, or common sense in this problem of test weighting. The point which we wish to make is that rational analysis and judgment are the *fundamental basis* for the use of tests in personnel selection, and that statistical procedures are of value as a guide and supplement to, and not as a substitute for, professional judgment.

Statistical methods will tell us how well each test predicts a particular criterion measure, and how tests should be weighted to give the best prediction of that measure. They will not tell us in any definitive way whether that criterion measure is a relevant or complete measure of the behaviors which represent success on that job. That is a matter of judgment. Statistics

will not tell us which of several criterion measures is the most important and should count most as we sum up our available data. That is a matter of judgment. Statistics will not tell us what traits are important for success in the job as a whole but are not represented in our particular criterion measures. That is a matter of judgment. And a matter of judgment, too, is the decision on how implicitly to follow the statistical data with regard to validities and regression weights. If we judge that the criterion data which we have are the best that can be obtained and that our professional enterprise will be best served by following the empirical data exactly, we may have judged wisely but we have still judged. In another situation the sound judgment may be that the knowledge of the job represented in our job analysis is far superior, as a guide to test construction, to any empirical validation which it is feasible to obtain for that job, and that we should rely heavily upon our rational analysis in selecting and weighting tests.

Judgment necessarily enters in at some point in any personnel selection enterprise. It may enter early in the scheme of things, when a criterion measure is selected and the decision is made to concentrate explicitly on predicting that criterion. It may enter also at the final stage of planning a test battery, when tests are weighted or test weights are modified from those specified by the statistical analysis because of knowledge of aspects of the job other than those represented in the criterion measure. The use of judgment is no more reprehensible in the latter case than in the former. Rather, it is the blind use of statistics without judgment which is to be condemned.

### THREE MAJOR TYPES OF TESTING PROGRAMS

In the use of tests to evaluate personnel for job assignment we can recognize three major patterns. These are as follows:

1. The use of tests as a screening device to qualify personnel for assignment to a single job or type of training. (Selection.)
2. The use of tests as a multiple screening device to qualify personnel simultaneously for assignment to one or more of a number of jobs or types of training. (Multiple selection.)



3. The use of tests to determine to *which one* of a number of jobs or types of training each person should be assigned. (Classification.)

In each of these patterns we must recognize a somewhat different purpose underlying the testing. Each of these purposes brings forth certain distinctive problems and calls for somewhat different emphases in test development and test analysis procedures. We shall need to consider each in turn.

### *Use of a test battery for selection*

In simple selection, we are dealing with a testing program which is being developed with the single purpose of picking personnel for a single job or training program. Each individual is being considered only as an applicant for that one job category. For each applicant, the problem is merely whether to accept or reject him for that one position. An illustration of a program of simple selection is that undertaken by a medical school or law school in screening out the most promising from among a large number of applicants for admission. Another illustration is a program being carried out by a life insurance company to select salesmen. The one concern in these programs is to select the individuals who will do best in that one job.

The simple selection problem is the one for which our statistical procedures of test analysis are best fitted. When a battery of tests is being developed to pick personnel for assignment to a single job, the simple purpose is to get a score with the highest possible validity in terms of an adequate criterion of success on that job. (We must, of course, recognize practical factors of economy, convenience, etc., but these cannot enter into the statistical analysis of test results. They are supplementary non-statistical considerations which must be used to guide the analysis and interpretation of the test results.) Test development activities are directed at the single purpose of getting the most accurate prediction of success in that one job. Working in the framework of multiple correlation, the single purpose is to make the multiple correlation between composite score derived from the test battery and criterion of job success the maximum.

With the simple goal of predicting success in a single job, the natural selection approach is to see how well the individual



actually does in a sample of the job duties. The applying typist may be given a standardized typing test; the machinist may be set to work on a lathe with materials, tools, and a blue print. The approach through a single direct test of proficiency in the job may be the best in certain cases, but in most instances a proficiency test must be supplemented, and often it does not represent a practical undertaking. The usefulness of proficiency tests is limited by several factors:

1. Proficiency tests may call for equipment, such as milling machines, airplanes, etc., which are prohibitively expensive for use in a personnel selection program.

2. Proficiency tests may require too large an expenditure of testing time in order to get a reasonably reliable sample of the individual's performance on the complex job task.

3. The job duties may be of such a subtle, varied, and intangible nature that it is impossible to reproduce them in a test situation. Thus, the varied adaptations of a salesman to a customer or the many little tasks of a private secretary may not lend themselves to testing.

4. The applicant may yet have to learn the task, so that it may be impossible to test his proficiency at the present time. Thus, applicants for pilot training could hardly be turned loose in an airplane, nor could potential engineering students be expected to design a bridge.

These conditions make a place for aptitude tests in personnel selection—tests which are not the job, but which are prognostic of future success in it.

Even when direct testing of proficiency is not feasible, development of aptitude tests for the selection of personnel for a particular job tends naturally toward the development of tests that closely resemble the criterion task both in content and in complexity. One can readily see the rationale for having the content of the testing situation resemble as nearly as possible the actual duties on the job. One reasons that, the more nearly the test approaches the job or some phase of the job, the more accurately test performance will predict job performance. Thus, for pilots one constructs motor coordination tests which use an airplane-type stick and rudder; for navigators one constructs a table-reading test, using data on drift, air speed, and the like;

for medical school students one constructs a test of ability to read and comprehend prose passages dealing with medical content. Each test is planned so that it measures certain general functions of human behavior, but also so that it measures those functions with the specific materials and in the specific situations that are likely to occur in the job in question. In this way, one hopes that factors of specific content as well as factors of general function may give the test validity for the job being studied.

A related but somewhat different tendency in the construction of tests for a single job is the tendency to make the tests complex. A job is ordinarily complex, requiring the individual to do a number of different things, often at the same time. In the effort to reproduce these complexities the test is likely also to become complex. This was illustrated in the program of wartime test development in the AAF by the large number of complex pursuit and coordination tests which were developed for pilot selection, requiring the individual simultaneously to use a number of controls and to respond to a variety of signals and cues. The defense offered for complex tests based on the materials of the job is that as individual tests they tend to have relatively high validity for the job for which they were particularly tailored. This seems often to be true. It is further urged that the use of material relating to the specific task on the one hand and the introduction of complexity of function on the other introduce valid variance not covered by *any number* of more analytical tests of simple mental functions formulated in terms of more general materials. This may also be true, though it is a point that would be very difficult to demonstrate conclusively. It did seem true in the AAF program for pilot and navigator selection that the validity achieved by complex tests was greater than could have been produced by any combination of simple and relatively "pure" tests available at the time.

When the personnel psychologist is concerned with a pure selection problem, the one criticism that can be leveled at the complex type of test, developed solely with an eye to its validity and without regard to its intercorrelations, is that each test of this type tends to have relatively high correlations with all other tests developed for that job in the same way. Since each test

involves a complex weighting of a number of the factors that enter into success on the job being studied, each tends to overlap markedly with every other. Furthermore, in so far as each different test is built with the same concepts of the job in mind, the tests overlap not only in their valid variance, but also in their invalid variance. The errors in interpreting the job also are generally perpetuated. The high intercorrelations indicate that relatively little gain will result from adding to an existing test or battery of tests of this type other tests of the same sort. The validity resulting from two or three tests is near the maximum that can be achieved. Whether the final multiple correlation will be higher for a test program based on complex job-analogy tests with relatively high validities and high intercorrelations or for a battery of relatively pure tests of simple psychological functions with lower validities and lower intercorrelations remains an open question.

#### *Use of a test battery for multiple selection*

The term *multiple selection* designates a set of operations through which an individual is simultaneously determined to be qualified or not qualified for each one of a number of job specialties. In the air-crew selection program of the AAF, for example, a test battery was administered to each man and a score was obtained for bombardier aptitude, navigator aptitude, and pilot aptitude. At any given time minimum standards were in effect for each of the air-crew specialties, and by comparing the aptitude scores of a given man with the standards it was possible to determine for which one, two, or three specialties he was qualified. Subsequent classification and assignment was to one of the specialties for which he was qualified. The process of qualifying for a job category is the same as it was for the simple selection described in the previous section. However, it is carried out for several job categories at the same time.

In using tests for purposes of multiple screening, the theoretical situation is not essentially different from that of simple selection. In a sense two, three, or more testing batteries are all given at the same time, each of which gives rise to a set of regression weights for predicting success in a particular job category. Each

battery can, in theory, be developed in complete independence of the others and carried to the point of giving the best possible prediction of criteria of success in the particular job. In practice, however, the need for efficient use of limited testing time precludes the development of such parallel independent batteries. It might be possible to proceed in that way if the number of job specialties were only two or three, but the approach would become hopelessly inefficient, unwieldy, and time-consuming with a larger number of job categories. It then becomes necessary to use each test for predicting success in several jobs.

As it becomes necessary to use a single test in the prediction of success in not one but several jobs, it obviously becomes less defensible to design the test in terms of the duties of a particular job. Of course, the test may still be conceived as functioning *primarily* for a single job, being used incidentally in the prediction of success in other job specialties in so far as it is found empirically to be predictive of success in those job specialties. Thus, in the battery of AAF air-crew tests certain tests were conceived primarily as pilot tests, others as navigator tests, and still others as bombardier tests. However, each test was weighted for any air-crew specialty for which analysis of validities and intercorrelations indicated that it should receive weight. Furthermore, in expanding the use of the battery to additional job categories, the tests developed for pilot, navigator, and bombardier provided the basic battery for the new specialty. Thus, these tests were re-weighted for flight engineers when it became necessary to select men for that assignment. However, this procedure was considered a temporary expedient pending the development and validation of tests more specifically directed at predicting flight engineer criteria. However, the procedure of using more or less distinct sub-batteries was possible only because an extensive battery of some twenty tests was being used. When selection is being carried out for several jobs at the same time, it is usually important to pick tests in terms of their effectiveness for more than a single job category.

If tests are to be designed less in terms of the activities of a particular job, they must be designed more in terms of general categories of human behavior. The approach to test development in terms of aspects of human behavior starts off with the



search for and definition of behavior categories. Categories may be drawn to a large extent ready-made from the language of the introductory psychology textbook or of everyday speech. In this way one may set out to build tests of "judgment," "attention," "observation," "memory," and the like. However, such verbal labels provide only starting points for test construction, and as the necessary steps are taken to translate the categories into usable tests certain difficulties are likely to arise. In particular one is likely to find that tests which purport to measure essentially the same category of behavior, and which should be functionally nearly the same, show only moderate correlation, and tests which purport to measure different categories, and which might therefore be expected to be independent, are in fact correlated to a fairly substantial degree. In other words, test scores often do not organize themselves into sharply defined clusters corresponding to a priori categories.

The failure of the pattern of test intercorrelations to confirm a priori categories has led psychologists to try to refine categories and revise tests in the light of the test intercorrelations. The effort to develop refined and more useful categories depends in every case on obtaining the matrix of test intercorrelations. There is great diversity, however, in what use is then made of the intercorrelations by different workers. On the one hand, the intercorrelations may serve primarily as material for sophisticated inspection, in terms of which the tests are re-interpreted and hypotheses are formulated as to new test operations which are expected to provide more nearly unique and uncorrelated tests. Thus, examination of the intercorrelations of a test which was designed to test perceptual speed may show it to be correlated with several tests of a numerical character. This may lead to the revision of the perceptual test and the exclusion from it of all numerical content, with the hope that the change will yield a purer test with lower intercorrelations. Tests revised in this way will, it is hoped, define more nearly separate and distinct dimensions of human behavior. On the other hand, the same goal is sought through the complex series of operations involved in factor analysis. Factor analysis undertakes to resolve the test intercorrelations into a number of independent components, and



through rotations of these components to identify each with both some nameable aspect of behavior and some test or tests.<sup>8</sup>

The goal in the refinement of categories is a set of categories which are mutually independent and collectively inclusive. As far as this goal can be achieved the resulting set of categories and the set of measures to represent them will have both logical and practical advantages. It is simpler to think and talk about a set of categories all of which are separate and distinct rather than interrelated in varying degrees. From the practical point of view, independence of the several tests will contribute to the efficiency of the battery as it is used for multiple selection. Each test will measure a new aspect of human behavior, with a minimum of duplication of what has been covered in other tests. A maximum scope of human behavior will be evaluated within a given period of testing time. In so far as predictions must be made for a number of jobs involving a variety of types of duties and in so far as it is consequently necessary to evaluate many different aspects of behavior, this non-overlapping in different tests may be a matter of great practical importance. It becomes a matter of theoretical importance in connection with the problem of classification which we shall consider next.

### *Use of a test battery for classification*

In multiple selection the goal remains the relatively straightforward one of obtaining the maximum accuracy in the prediction of success in each one of several job specialties taken singly. Since it is not practical to design a separate battery for each job specialty, compromise with the ideal must be made, and some loss in the accuracy of prediction of single job categories is tolerated in order that the prediction of others may be improved. The practical goal is that the average validity for all job specialties, with appropriate weight being given for the importance of

<sup>8</sup> An introduction to the methods of factor analysis is given in J. P. Guilford, *Psychometric Methods*, McGraw-Hill Book Co., New York, 1936, Chapter 14. A more complete presentation of the whole topic may be found in Godfrey H. Thomson, *The Factorial Analysis of Human Ability*, Houghton Mifflin, Boston, 1939, and L. L. Thurstone, *Multiple-Factor Analysis*, University of Chicago Press, Chicago, 1947.

each job, be a maximum within the limits of the time and facilities available for testing.

As soon as the task becomes one of classification, an entirely new element is introduced into the goal of testing. We define a strictly *classification* program as one in which *each man must be used* in some one of the available job specialties. The purpose of testing becomes to determine his *relative fitness* for each of the different duties. Final classification must take account not only of the single individual's relative promise in the different job categories, but also of the number of vacancies to be filled in each category and the relative fitness of the other candidates.

A pure problem in classification arises when we have  $N$  individuals to be assigned to  $N$  positions in  $k$  different categories. The goal is to contribute the maximum to the over-all effectiveness of the organization as a result of the assignments. In this situation there are no more men than jobs, so it is not possible to reject any man completely. For this problem, it is possible to make only limited use of the *absolute level* of the individual's aptitude for or evidences of ability in any particular job. The critical factor will be *differences in level* of aptitude for or ability in the  $k$  possible assignments. The situation is further complicated, in the practical case, by the fact that there are frequently certain job categories in which it is more important to have the maximum level of effectiveness than others. Thus, in classifying ground personnel for an air force, it may be more important to have the best possible radar maintenance men than to have the best possible cooks. A manufacturing concern may well be more concerned about the efficiency of milling machine operators than the efficiency of the individuals who clean the factory.

The classification problem has been rather generally formulated in the preceding paragraph. We shall now try to make a more formal statement of it, indicating the points which must be covered if an analytical, mathematical attack on the problem is to be made.

*Given:*

1. A limited number  $N$  of individuals available for job assignment.

2. The same number  $N$  of jobs to be filled, these jobs being of  $k$  different kinds ( $k \leq N$ ).
3. A set of measures of different aspects of individual aptitude or achievement.
4. Data on the validity of each of the measures in item 3 for each of the job categories in item 2, together with the correlations among the predictors and the correlations among the criterion measures.
5. Weights for each job specialty, indicating the importance attributed to having individuals with the highest aptitude for that job assigned to it.

*Required:*

A procedure for assigning the complete group of individuals in such a way that the weighted sum (by weights in item 5 above) of the aptitudes of all the men in all the jobs shall be the maximum.

The above statement presents the classification problem in its pure form. In practice a testing program is often used jointly for purposes of multiple selection and of classification. That is, it may be used both to disqualify a certain number of the inept and to make positive classification of those who qualify for one or more job categories. The problem of classification is clearly a complex one, and an analytical mathematical approach to it is very difficult even when satisfactory values can be established for all the "givens" listed above.

In a classification program we are no longer primarily interested in the *level of aptitude* for single jobs, since it is specified that we must use even the least able men somewhere. We are now interested in *differences in aptitude* for different jobs. It is no longer sufficient to predict success in job A accurately and to predict success in job B accurately; we must predict accurately difference in success between job A and job B. This means that we must be interested not so much in the validity of our test or our composite score for job A and its validity for job B, but rather in the extent to which it leads to differential predictions for the two jobs.

A reasonably straightforward mathematical approach can be made to the problem of classification when the number of job

categories is only two. This will be developed here in part for its own sake and in part to clarify the nature of the general problem. When the number of job categories becomes more than two, however, some entirely new method will have to be developed for the mathematical solution of the classification problem.

When there are only two job categories, A and B, we can make our best prediction of each, in the least squares sense. These predictions will be expressed as

$$\tilde{y}_A = {}_A\beta_1x_1 + {}_A\beta_2x_2 + \cdots + {}_A\beta_kx_k \quad (13)$$

$$\tilde{y}_B = {}_B\beta_1x_1 + {}_B\beta_2x_2 + \cdots + {}_B\beta_kx_k$$

where  ${}_A\beta_k$  = weight to be applied to variable  $k$  in predicting success in job A.

${}_B\beta_k$  = weight to be applied to variable  $k$  in predicting success in job B.

There now exists, for each individual, a single value which represents the difference between his predicted success in job A and his predicted success in job B:

$$\tilde{\Delta} = (\tilde{y}_A - \tilde{y}_B) \quad (14)$$

For simplicity, both jobs are here considered to be equally important. However, a weighted difference could replace this simple difference if we desired to attach more weight to placing promising applicants in one job category than to placing them in the other. For classification, we can use the value,  $\tilde{\Delta}$  which represents the predicted difference in success in the two jobs. Individuals can be ranked in order by the value of  $\tilde{\Delta}$ . The number required for assignment to job A may be selected from the end of the ranking which represents greater likelihood of success in job A. Referring to equations 13 and performing the subtraction, we have

$$\begin{aligned} \tilde{\Delta} = \tilde{y}_A - \tilde{y}_B = ({}_A\beta_1 - {}_B\beta_1)x_1 + ({}_A\beta_2 - {}_B\beta_2)x_2 + \cdots \\ + ({}_A\beta_k - {}_B\beta_k)x_k \end{aligned} \quad (15)$$

The weight for each variable for predicting the difference in aptitude for the two jobs is the difference between the weights for the separate jobs.

The validity of the estimate,  $\tilde{\Delta}$ , of differences in aptitude can be expressed by the familiar formula for the correlation of sums and differences. Let

$A$  represent score predicting success in job  $A$ .

$\alpha$  represent actual success in job  $A$ .

$B$  represent score predicting success in job  $B$ .

$\beta$  represent actual success in job  $B$ .

Then  $(A - B)$  is the predicted difference in success on the two jobs and  $(\alpha - \beta)$  is the actual difference in success. The validity of the differential prediction is given by the formula

$$r_{(A-B)(\alpha-\beta)} = \frac{(r_{A\alpha} - r_{A\beta}) + (r_{B\beta} - r_{B\alpha})}{\sqrt{1 - r_{AB}} \sqrt{1 - r_{\alpha\beta}}} \quad (16)$$

An examination of this formula brings out two interesting points. In the first place, it can be seen that the most critical factor in differential prediction is the difference between the validity of a composite score for the criterion which it has been developed to predict and its validity for the other criterion. This is brought out by the numerator of the above expression. The actual *level* of the validity coefficients enters into the result only indirectly; it is the size of the difference between them that is of direct importance. Our goal must be to develop tests that will make these differences a maximum.

The second point to be noted is that for a specified set of validity coefficients high correlation between the prediction scores (and between the criterion measures) makes for more rather than less validity of differential prediction. This matter of intercorrelation of prediction scores is not likely to enter into the plans for test development. Attention is likely to be centered on the validities of each test for the two job criteria. However, it is interesting to note, with an eye to the problem of classifying among three job categories,  $A$ ,  $B$ , and  $C$ , that, when it is necessary to use tests of the same function in the batteries for  $B$  and  $C$ , in order to differentiate these jobs from  $A$ , it is preferable to use the *identical* score in both batteries. The common errors of measurement, which tend to increase the correlation between the predictors of  $B$  and  $C$ , will be an advantage rather than the



reverse. In using scores for classification purposes, it is desirable that the errors of measurement and other completely non-functional variance in the several score composites be as *highly correlated as possible*. This non-functional variance tends thereby to be held constant for the different job categories and to be more or less partialled out of any difference score.

The discussion in the previous paragraphs indicates that for classification between two job categories we can use the single difference score which represents the difference between predicted success in one job and predicted success in the other job. With three job categories, however, there are three differences between pairs of scores predicting success in the single jobs, with four jobs there are six difference scores, and so forth. There is no simple way of simultaneously examining those several differences for all the applicants and deciding which assignment is most advantageous for each one. So far as the author is aware, this is still an unsolved problem.

Returning to procedures for test construction, we have seen that for a *classification* program the measure of the effectiveness of a test battery lies in the differential validity of the several prediction scores for the several jobs. A number of composite prediction scores can have different patterns of validity for the several criteria only in so far as they measure different behavior functions. The only validity of the battery for classification purposes, therefore, lies in the difference in function measured by the different scores. A test is of value in so far as it permits the differentiation of some function from other functions.

Let us suppose that we have a test which measures a pure trait of behavior, the trait being isolated by the best available statistical techniques and professional insight so that it is as completely unrelated to others and as psychologically meaningful as may be. We may expect the validity of this test to differ sharply from one job to another, since the trait is likely to be important for some jobs but of little or no importance for others. The differences in validity will not be blurred because other traits enter into the test score, other traits that are valid for different sets of jobs. A set of tests of this sort permits the maximum of difference in validity in composite scores for different job specialties, since the tests chosen for weighting for a particular job will include only those

factors that have validity for the job. They will not carry with them secondary sources of variance which have validity not for the job in which success is being predicted but for some other job or jobs. A composite score will have validity for jobs other than the one it is supposed to predict only in so far as those jobs actually call for the same traits, and not because a particular test measures some traits which are valid for each of the jobs.

In the case of highly complex tests, the situation tends to be reversed. The complex test is almost certain to reflect a number of traits or factors. This multiplicity of factors probably includes some that are valid for one job and some for another. Furthermore, other complex tests will often overlap the first one and measure a number of the same factors. A composite score built up of several such tests may represent an even more complex array of factors. This complex of factors may include a number of factors that have validity for the job on which the composite score is designed to predict success, and so the composite score may be quite a valid predictor of success in that job. However, the composite score is also likely to include a number of factors that are valid for other jobs, and so it may be fairly valid for these jobs too. Thus, the value of the score as an instrument for *differential* assignment is reduced.

It seems, in general, that it is for the task of classification that pure tests of single functions of human behavior have their greatest justification. For differential prediction, purity rather than high validity seems likely to yield the most effective test battery. It is in this type of program that the research devoted to the design of pure tests of distinct factors in human behavior is most likely to be fruitful.

## *The Analysis and Selection of Test Items*

The typical test for personnel selection is composed of a fairly large number of separate items. Each item requires the subject to exhibit certain knowledges or skills, or, in certain types of testing instruments, to state certain facts about himself or to express his attitudes on some point. In using test results for selection and in selection research we ordinarily extract a single score from the group of items which constitute a test or a part of a test. We combine the individual item responses in a single score, sometimes purely because it is a practical necessity to do so but sometimes because in addition we consider the separate items sufficiently uniform in character reasonably to be combined into a single score.

### FACTORS IN ITEM EVALUATION

The effectiveness of a test depends on the characteristics of the items which compose it. In both its reliability and its validity a test score is the resultant of the validities, reliabilities, and intercorrelations of its component items. In order to produce the most effective test, therefore, we must study each one of the pool of items from which the test is to be assembled. The choice of items for the final form of a test is based in part on the detailed specifications for the content of the final test which were prepared as a part of the process of planning the test. It is based in part on certain statistical characteristics of each item. There are two statistical aspects of the individual item with which we shall be concerned. The first is the difficulty level of the item for the group being studied. The second is the degree to which

the item differentiates those who are high from those who are low on some standard. This standard may sometimes be performance on the complete pool of items, in which case we are concerned with the internal consistency of the items. The standard may sometimes be an external criterion of job performance, in which case we are concerned with the validity of each individual item.

### *Item difficulty*

Let us consider first the role of difficulty as a factor in item selection. If an item is to be useful in distinguishing between those who are high and those who are low on a certain trait, it is apparent that the item must not be so easy as to be passed by every member of the group nor so difficult as to be failed by every member. In neither of these extreme cases does the item make *any* contribution to the discrimination which the test is to make between different individuals. Within these two extremes, the difficulty of a test item is still a significant factor in evaluating the item. In general a single item will make the largest number of discriminations between pairs of individuals if it is at a difficulty level such that it is passed by 50 per cent of the group.

Taking a single item, we can visualize the number of discriminations which it makes in the following way. An item which is given to 100 subjects and passed by one discriminates between that one and any one of the remaining 99, and thus makes 99 discriminations. An item which is passed by 10 and failed by 90 discriminates between each one of the group of 10 and each one of the group of 90. If we took the individuals by pairs, there would be  $10 \times 90 = 900$  combinations in which that item discriminated between the two members of a pair. An item which was passed by 50 individuals of the group of 100 and failed by the remaining 50 would discriminate each member of the first 50 from each member of the remaining 50. There would then be  $50 \times 50 = 2500$  combinations in which the item would discriminate between the members of a pair of individuals chosen from the total group of 100. Figure 1 shows the relationship between difficulty level, expressed as percentage succeeding with the item, and number of discriminations made by the item.

Clearly, the item that is passed by approximately one-half the individuals makes discriminations between many more pairs of individuals than the item that is passed by a very small or a very large percentage of the total group. Differences in difficulty do not affect the frequency of discrimination so much in the middle difficulty ranges (from 25 to 75 per cent success, let us say), but

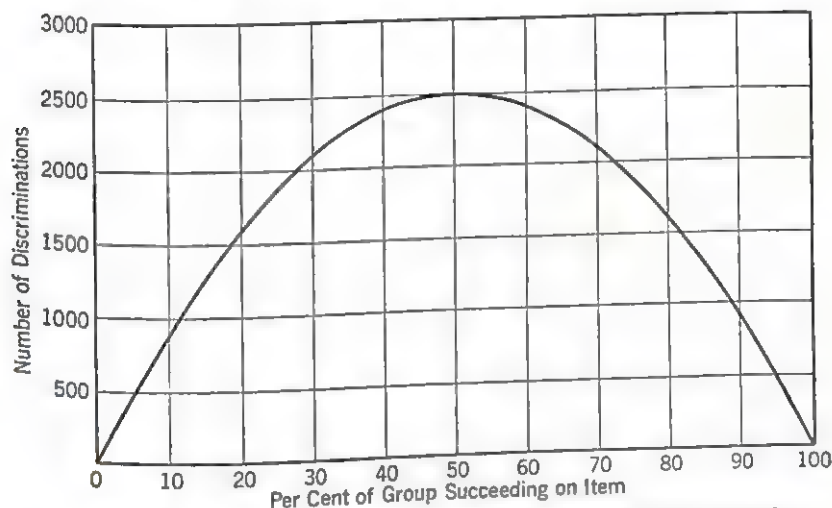


FIGURE 1. Number of discriminations as a function of item difficulty.

they become increasingly critical as the extreme values are approached.

Both theoretical and empirical studies indicate that a test discriminates most reliably among the members of a group being tested when the average difficulty level of the items in the test is such that approximately 50 per cent of the group succeed with an item. The exact distribution of item difficulties that will be most effective in any given case involves also a consideration of the relationship among the items. In proportion as the correlation between each item and each other item is high, a test will require a relatively wide scatter of item difficulties around the average level of 50 per cent difficulty. In proportion as the correlation between separate items is low, the most advantageous scatter of item difficulties may be expected to be small. This relationship is one which has been made clear only in a general sense and not expressed in a clear-cut analytical fashion. It is



not possible at the present time to specify for any given test with an obtained set of item intercorrelations what the optimum distribution of item difficulties should be. However, there is general agreement that what is required is a set of items which average 50 per cent difficulty for the group to be tested and which show some degree of scatter away from that value.

### *Item discrimination*

We turn now to a consideration of the ability to discriminate high-scoring individuals as an item characteristic. As was indicated, this discrimination may be in terms of total score on the test, or it may be in terms of some external criterion score of job performance. In some cases one type of analysis will be appropriate, in some cases the other. There may be some cases in which both seem reasonable. To clarify the situations that call for analysis of the relationship of item to total test score (internal consistency analysis) and the situations that call for analysis of the relationship of item to an external criterion (item validation), let us consider two types of testing instrument.

In the first type of testing instrument, the items in a given test blank are included in a single score purely from considerations of convenience and not because of an essential unity in the items themselves. This type of instrument is perhaps best illustrated by the biographical data blank, in which a great variety of questions are asked about the individual's past history. Certain personality questionnaires and interest inventories also fall into this class. As these instruments are tried out in preliminary form, they include a wide range of items which have only a superficial relationship to one another. They are held together not by any basic functional relationship, but rather by the superficial fact that all the items refer to facts of personal history, statements about likes and dislikes, or statements of individual feelings or behavior. They may have very little unity in terms of what the individual items purport to measure. Convenience and a certain superficial resemblance in the form of the items make it desirable to include them within a single test blank. There is no reason however to think of the items as representing a single unified aspect of the individual's development.

In such a situation, it is very appropriate to inquire into the validity of each single item. Each item is in a very real sense a little test all by itself. Each item must necessarily be judged on its own merit as far as validity is concerned. There is no determinable a priori basis for scoring the items and combining them into a score. The only basis for arriving at a selection of items or a scoring of items is in terms of the empirical validity of each item. In this type of test, therefore, it is entirely appropriate to determine the effectiveness of each item (and of each possible response to it) in discriminating with respect to some outside criterion of success on the job. An item response is valid in so far as the individuals who choose that response are more successful on the job, or more frequently successful, than those who do not choose that response to that item.

Where the separate test items represent such a heterogeneous assortment of materials, there is no a priori reason for expecting correlation among the separate items and consequently no justification for analyzing the relationship of a single item to a score based on the whole pool of items. As a matter of fact, as we noted, there is no basis for arriving a priori at a pooled score for a set of items, and consequently there is no meaningful score against which the internal consistency of the test items could be checked. In this case, then, there is no point in carrying out an internal consistency item analysis.

Let us consider, in sharp contrast with the type of test blank just described, a test which has been developed for the purpose of measuring a limited and specific segment of ability. This might, for example, be a test of the extent of general vocabulary. In such a test, each item purports to help assess some single unified aspect of ability. Furthermore by the nature of the test, expert opinion is able to identify in advance the "right" answer for each item. It is possible to score the items in the test without reference to an external standard. If a hundred such items have been prepared, the scoring key for the total test can be developed a priori without reference to any empirical results. In this case it is appropriate to ask to what extent a particular test item measures the same function in the individuals tested as is meas-

ured by the test as a whole. This is possible because (1) a total test score can be obtained and (2) such a score will be meaningful because it was the purpose of the test-maker that each item contribute to the measurement of a single homogeneous function.

When the items in a test are completely homogeneous in that each item measures exactly the same factors or aspects of individual ability in the same combination, correlation of the separate items with an external criterion of success loses its meaning. In so far as the single items measure exactly the same combination of abilities, they necessarily differ in validity only as one item is more reliable than another. That is, their correlations with an external criterion depend only on their reliabilities. When the items are truly homogeneous, a difference in item reliability is directly reflected in differences in correlation with total test score. In this case, all the relevant information about the item is provided by its internal consistency item analysis, and any further analysis of individual items against an external criterion is futile.

The two examples described in the previous paragraphs represent extremes. At the one extreme, we had a testing instrument in which each component item could be thought of as a separate little test by itself, having no intrinsic relationship to the other items and necessarily being validated on its own merits. At the other extreme we pictured a test composed of items all of which were homogeneous measures of exactly the same functions in an individual. In the first test only item validation makes sense as a procedure for analyzing single items. In the second only internal consistency analysis provides an appropriate measure for evaluating single items. In actual practice the situation may not be so clear-cut. Many tests may have a degree of both homogeneity and heterogeneity. It may be possible to score the items *a priori* and to arrive at a total score, and yet the functions measured by the items may have enough diversity so that validation of the single items is appropriate. It is significant to note, however, that, as validation of the single items becomes more meaningful, internal consistency analysis becomes less meaningful, and the more appropriate internal consistency analysis is, the less we may expect item validation to contribute.

## PROCEDURES FOR CALCULATING ITEM INDICES

*Indices of item difficulty*

An indication of the difficulty level of an item is given by the percentage of individuals in a given population who can answer the question or solve the problem. The smaller the percentage succeeding on the item, the more difficult the item, and vice versa. This type of index has been criticized on the grounds that infrequency of answering may sometimes be a function of the obscure and esoteric nature of the information called for by the item rather than any fundamental intellectual complexity of the concept or process required. Thus, there is no more fundamental intellectual difficulty in knowing that a wombat is an animal than in knowing that a dog is. However, knowledge of the meaning of wombat is certainly very much rarer than knowledge of the meaning of dog. "Difficulty" in this instance does not represent complexity of concept but rather rarity of the concept in common experience. Recognizing this limitation, we may still accept the percentage of those knowing the answer to an item as a useful operational measure of its difficulty.

The raw percentage succeeding on an item gives a crude difficulty index. However, that percentage suffers from two limitations. In the first place, some of those who got the correct answer on the item may have done so by chance, that is, by guessing or by some other procedure which did not involve knowledge of the required information or working through of the problem. In the second place, it is only for a rectangular distribution of ability that percentage of success provides a linear scale of difficulty. If we consider the ability measured by the test as having approximately a normal distribution, equal differences in percentage of success will generally not correspond to equal steps in the scale of difficulty. We shall now consider the adjustments and transformations which may be applied to take care of these problems.

First, let us consider the problem of correcting difficulty indices for chance success. Our effort here is to get an estimate of the percentage of individuals in the group who arrived at the right answer through correct knowledge or correct reasoning and

to rule out those who got the answer by guesswork. Any correction for guessing involves something of an approximation. The usual allowance is based upon two assumptions. In the first place, it is assumed that an individual who selects an incorrect answer to a test item does so on the basis of absence of information or understanding rather than on the basis of misinformation or misunderstanding. In the second place, it is assumed that for a person who does not know the correct answer all the response options on an item are equally attractive. If we accept these assumptions, we shall expect a certain percentage of those who do not *know* the right answer to select that answer by chance. This percentage will be the average of the percentages choosing each of the wrong answers. We subtract this percentage from the raw percentage of success on the item to arrive at our estimate of the percentage actually knowing the answer to the item.

The correction which we have just described is given in algebraic form by the formula

$$P_c = \frac{R - \frac{W}{n-1}}{R + W + O} \quad (1)$$

in which  $P_c$  = the percentage actually knowing the answer to the item.

$R$  = the number giving the right answer.

$W$  = the number giving a wrong answer.

$O$  = the number choosing to omit the item.

$n$  = the number of response alternatives for the item.

Of the above concepts, the only one that is likely to present any ambiguity is the number choosing to omit the item. An omitted item is one that was read by the examinee and then omitted by choice. In a speeded test there may be some confusion between items that were omitted by choice of the examinee because he did not know the answer to them and items that the examinee did not have time to attempt. A practical basis for differentiating between the two is to assume that all items up to and including the last one for which an answer was marked were attempted and that the first unanswered item after the last answered one



was also attempted. We then assume that starting with the second item after the last one answered the remaining items were not tried. If, for example, on a 100-item test the last item for which an answer was marked was the 84th, all unmarked items up to and including No. 85 would be counted as omissions, and Nos. 86-100 would be counted as not attempted.

It should be noted that the formula in the preceding paragraph depends on the two assumptions stated above. The two assumptions will rarely be exactly satisfied. Deviation from the assumed conditions will have opposite and somewhat compensatory effects in the two cases. If one or more of the misleads for an item are quite implausible and unattractive, the number of effective response alternatives for that item is reduced. This means that the probability of obtaining the right answer by guessing is increased since the guess is between, say, three instead of five alternatives. In this case, the formula would tend to under-correct for the possibility of guessing the right answer. On the other hand, individuals often arrive at a wrong answer not as a result of ignorance and blind guessing but through wrong information or wrong mental processes. If all the individuals who gave wrong answers gave them for genuine reasons, guessing would not enter into the picture. In this case, all those who gave the right answer must also have given it because of correct information and correct mental operations. No correction for chance would then be indicated, and our formula would be over-correcting. In practice we do not know to what extent these two compensatory factors are operating, so that the corrected value for the percentage succeeding on an item is only an estimate of the percentage in the group who really knew the answer or solved the problem.

Though the corrected value for the percentage knowing an item is only an approximation and is limited by the underlying assumptions which we have discussed, it may usually be expected to give a truer picture than the raw value. It is of value particularly for comparing the difficulties of items which either (1) differ sharply in the frequency with which examinees omit the item or (2) differ in the number of the misleads. Raw percentages of correct response cannot meaningfully be compared in these cases.

The inequality of units which results from using percentage

of success as an index of item difficulty can be overcome if we are willing to make some assumption as to the shape of the distribution of the trait being tested in the group to whom the tests were administered. The usual assumption is that the trait is normally distributed. If that assumption is made, the percentages can be translated into scale values on the base line of the normal curve. Thus, 50 per cent success corresponds to a scale value of 0.00, 75 per cent success to a scale value of  $-0.67$ , 25 per cent success to a scale value of  $+0.67$ , and so forth. With this type of scaling, it is possible to compare the difficulty of items which have been tried out on different groups, provided that there is some way to determine the difference in mean and standard deviation of ability in the two groups.<sup>1</sup>

### *Indices of item discrimination*

Many indices have been proposed to show the degree to which an item is effective in discriminating between those of high and low ability on either total test score or some outside criterion. Any attempt to list and discuss all the proposed types of statistical manipulation would be futile. In this section, therefore, a selection is made of those procedures which appear to have special advantages either because of extreme simplicity and ease of computation or because of the effectiveness with which they fit into the pattern of standard statistics for analyzing and combining test scores.

When we are carrying out an item analysis we encounter two major types of situation. In one, we are relating performance on the item to some type of continuous measure. This continuous variable is often score on the total test of which the item is a component, but the continuous measure may be some type of external criterion. The second situation is one in which we are relating performance on the item to some dichotomy, either an arbitrary dichotomy into which the continuous variable of total test score has been thrown or a dichotomy on some criterion variable. The procedures which are available to us in these

<sup>1</sup> The problem of appropriate indices of item difficulty is considered, together with other problems of item analysis, in F. B. Davis, *Item-Analysis Data*, Harvard Education Papers No. 2, Graduate School of Education, Harvard University, Cambridge, Mass., 1946.

two situations are somewhat different and need to be considered separately. We will turn our attention first to the comparison of item performance with score on some continuous variable.

#### ITEM ANALYSIS AGAINST A CONTINUOUS SCORE

When the variable against which an item is being analyzed is available to us as a continuously distributed score, there are several approaches for dealing with it. The most elegant procedure is to retain and use the continuously distributed score. However, as we shall see, the computational procedures involved in this case become quite laborious. The second alternative is to set some arbitrary dividing point on the scale of scores and throw the continuous variable into an arbitrary dichotomy. That is, we can split the group into the top half and the bottom half, for example, in terms of the continuous score variable. A third alternative is to compare two relatively extreme groups with regard to the continuous variable. That is, we may take a fraction from the top of the group and a fraction from the bottom of the group, rejecting a fraction from the middle, and compare performance on the single test item for those two extreme groups.

We shall first consider procedures for evaluating items when the continuous variable is retained as a quantitative score. When we wish to determine the relationship of each item to a continuously distributed score, the procedures available to us are the biserial correlation coefficient and the point biserial correlation coefficient, the nature of which were expounded in Chapter 6. In the present case, performance on the item constitutes the dichotomy. Either the individual gave a particular response to the item or he did not. Total test score constitutes the continuum. We wish to establish the degree of relationship between the single item and the total score. We may use either the biserial correlation coefficient or the point biserial, depending on whether we wish to consider the dichotomy on the item as real or as artificial.

If we think of success or failure on the item as representative of some continuous underlying ability and of the split into those who succeed and those who fail as an arbitrary splitting of individuals with regard to this underlying quality, the particular

location of the split depending on the difficulty of the particular item, then it seems reasonable to use a biserial correlation coefficient. That is, item success or failure is thought of as the symptom of some basic underlying variable. Interest is considered to be centered primarily in the underlying ability of which the item is only a symptom. The biserial correlation coefficient has one feature that is particularly worthy of consideration in this connection. As we noted in the previous chapter, the value which it takes does not depend on the proportion in the passing and in the failing group. This means that the index of item validity or discrimination is separated entirely from the matter of item difficulty. The two facts about the item can be separately determined, and in evaluating the item separate account can be taken of each.

The point biserial correlation differs from the biserial in the nature of the assumptions which it makes and in the effect which item difficulty has on the resulting values. The point biserial assumes in effect that those who pass and those who fail on a test item are two categorically distinct groups. It assumes that the groups can be distinguished but that within each group all the members are to be thought of as alike. The use of point biserial correlations in item analysis has the practical effect that the resulting indices are in part a function of item difficulty. Point biserial correlations tend to become smaller, other things being equal, as the proportion succeeding on the item departs from 50 per cent. In general neither of these characteristics of the point biserial recommend it for item analysis. The assumption that passers and failers on an item are categorically distinct groups is not a very reasonable one.<sup>2</sup> Certainly both those who pass and those who fail represent a range of ability with regard to the underlying function which the item to some degree measures. The confusion of item validity with item difficulty also appears generally undesirable. It seems much more defensible to obtain separate information on the validity or discriminating quality and on the difficulty of an item and then to combine those two facts explicitly and rationally rather than

<sup>2</sup> For a few items of the biographical data type the assumption of categorical differences seems appropriate. Items referring to sex, marital status, race, etc., are examples.

to use a single index which combines them in unspecified proportions.

The chief objection to the biserial correlation coefficient as an index of an item's ability to discriminate appears to lie in the direction of computational labor. The calculation of a biserial correlation coefficient for each of the often very substantial number of items in a test often represents a burden too great for the personnel psychologist to undertake. It is largely on this account that we shall consider other procedures for estimating the extent to which an item discriminates between the high- and the low-scoring individuals on a continuous measure. ✓

#### ITEM ANALYSIS AGAINST DICHOTOMIZED GROUPS

In all the other item analysis procedures discussed here the continuous nature of the variable against which the item is being compared is sacrificed in the interest of ease of computation. Two groups are specified, representing a higher and a lower group with respect to the continuous variable. In the procedures which we shall consider first, these two groups include between them all the cases. That is, some dividing line is arbitrarily set up for this continuous score; those falling above the dividing line constitute one group and those falling below the dividing line constitute the other. The usual and simplest procedure is to choose the dividing line so that the total group is split into two halves. Essentially the same procedures apply when the dichotomy is already established in the data, i.e., when the criterion is a dichotomy.

Where a group has been split in this manner, there are several possible approaches to item evaluation. One is extremely simple and serves quite adequately the needs of many makers of informal tests and many analysts interested only in some help in the editorial improvement and refinement of their tests. This simple procedure is merely to compute the percentage succeeding on the item in the upper group on total test score and the percentage succeeding on the item in the lower group on total test score. A crude evaluation of the item may be obtained from these two percentages. As a procedure for weeding out the most unsatisfactory items, this technique is reasonably satisfactory. It has the advantage of great simplicity and a minimum of computation.



All that is required is to separate the answer sheets of the subjects tested into two piles in terms of total score and to tally the number of correct responses, incorrect responses, omissions, and items not tried for each item within each fraction. Either the raw percentage of correct responses, or the percentage corrected for chance successes may be computed from these figures. The corrected percentages are obtained by applying formula 1 on page 234 to the count for each fraction of the group.

If further refinement is desired in the treatment of the percentages of success and failure on each item for the two fractions which comprise the complete group, it is possible to compute either a tetrachoric correlation or a phi coefficient from these percentages. These follow the procedures indicated in Chapter 6. The only difference is that here we are dealing with an item response rather than a separate dichotomous variable. We have a fourfold table based on the categories of success and failure on the item and high or low on total score. The percentages in each of the four cells of this table provide the basis for computing either the tetrachoric correlation or the phi coefficient. Which of the two we decide to use depends on the type of assumption made with regard to the continuity of the underlying distributions in the two variables. If we assume that they are continuous and normally distributed, the tetrachoric correlation is appropriate. If we assume that they represent discrete categories, the phi coefficient is chosen. The tetrachoric correlation is unaffected by item difficulty, whereas the phi coefficient, like the point biserial correlation, reflects the difficulty level of the particular item.

### *Item analysis with extreme groups*

If we decide to translate our continuous scores into the two categories of a higher and a lower group, we immediately introduce the possibility of analyzing only two extreme groups, dropping out those in the middle. If the relationship of item score to test score is linear, so that the percentage of success on the item increases steadily as total score increases, taking groups from the extremes will sharpen the differences which we observe in a single item. An item will clearly make a sharper distinction between the top tenth and the bottom tenth of the total popula-

tion than it will between the top half and the bottom half. However, this increasing sharpness of discrimination is more or less balanced by the loss of information which results from including only some of the available individuals. Though an item would discriminate more sharply between the top 10 and the bottom 10 of 1000 individuals than it would between the top and bottom 500, we may anticipate that a comparison of different items based on the top and bottom 10 of 1000 would be less dependable than the comparison based on the top and bottom 500. Is there an optimum point at which the ideal balance of sharpness of discrimination combined with stability in the resulting values is obtained? Kelley<sup>3</sup> has shown that the ratio of the obtained difference to its standard error is a maximum when the top group and bottom group each includes approximately 27 per cent of the total population tested. That is, given that we are going to sacrifice the quantitative nature of our test scores, we can get the most accurate arrangement of items in order from most to least discriminating if we base our item analysis on only the top and bottom 27 per cent of the total group. Using either a larger or smaller percentage than this results in a loss in the accuracy with which the items can be ranked from most to least discriminating. This appears to be one case in which we can both have our cake and eat it too. We can reject the middle 46 per cent of cases, reducing the clerical labor of tallying almost by half, and at the same time increase the precision of our results. In general, then, any program of item analysis which proposes to sacrifice the quantitative nature of a continuous test score in order to gain economy in computation should also exclude the middle 46 per cent of the group. The result will be both a substantial additional economy in tabulation and some increase in the precision of the results. Our next concern is to consider procedures for dealing with item statistics based on the top and bottom 27 per cent of the total group.

The first step in any evaluation of test items by comparing the top with the bottom 27 per cent of the total group tested is to determine the percentage of cases succeeding with the item in

<sup>3</sup> T. L. Kelley, "The Selection of Upper and Lower Groups for the Validation of Test Items," *J. Educ. Psychol.*, **30**, 17-24 (1939).

the top group and the percentage of cases succeeding with the item in the bottom group. The procedure of correcting for chance success is the same as it was when the results for all individuals were analyzed. Formula 1 on page 234 is also used, and it usually is desirable to make the correction if there are many omissions by the group taking the test.

The most satisfactory item validity index based on the upper and lower 27 per cent is the estimate of the coefficient of correlation between item and test obtainable from tables prepared by Flanagan.<sup>4</sup> These tables are based on the assumption that the variables underlying both item success and test score have a continuous normal distribution. Utilizing tables of the normal bivariate frequency distribution, Flanagan calculated correlation coefficients corresponding to the possible percentages of success in the upper and lower groups. These correlation coefficients are estimates of the product-moment correlation between the two underlying continuous normally distributed variables.

Flanagan's table, which is reproduced in Appendix B, makes it extremely simple to compute item validity coefficients from the percentages of success in the upper and lower 27 per cent. By entering the table in the appropriate row and column, the correlation may be read directly.

One limitation in the use of any correlation coefficient as a measure of degree of relationship between two variables is the fact that units on the scale of correlation values do not have the same significance as one goes from small to large correlation values. The change in correlation as one goes from 0.20 to 0.25 is not comparable to the change as one goes from 0.90 to 0.95. In an effort to compensate for this limitation in the units in which correlation coefficients are expressed, Davis<sup>5</sup> has recently developed an item index which is based on Fisher's  $z$ -transformation of the correlation coefficient. Flanagan's correlation values

<sup>4</sup> J. C. Flanagan, "General Considerations in the Selection of Test Items and a Short Method of Estimating the Product-Moment Coefficient from the Data at the Tails of the Distribution," *J. Educ. Psychol.*, **30**, 674-680 (1939).

<sup>5</sup> F. B. Davis, *Item-Analysis Data*, Harvard Education Papers No. 2, Graduate School of Education, Harvard University, Cambridge, Mass., 1946.

are converted into  $z$  values and multiplied by a constant such that the index becomes 100 when 99 per cent of one group succeeds with the item and only 1 per cent of the other group succeeds. This procedure yields indices in which the units are approximately equal throughout the scale. The indices also are convenient in that there are no decimal points to cause confusion or error. However, these indices are not readily related to the framework of statistical values with which most workers are familiar.

### *Item analysis against a criterion available only as a dichotomy*

The procedures for calculating the item indices discussed so far were developed with reference to a variable which yields continuous scores. It is sometimes necessary to relate item success to a dichotomous criterion of job performance. This situation arises when we are studying the validities of items with respect to some such criterion as graduation versus elimination in training. The statistical procedures available to us are essentially the same as those that we can apply to a continuous score when the continuous score has been forced into a dichotomy. We have discussed these in a previous section. The only difference is that in the present case the percentages falling in the "success" and "failure" categories are predetermined by the criterion dichotomy. These percentages will, of course, usually not be 50-50, and the computation of tetrachoric correlations or phi coefficients will have to take account of the percentage values.

## USE OF ITEM VALIDITIES

When the items in a test are more or less heterogeneous, so that different items may be expected to measure somewhat different combinations of abilities, it becomes appropriate to determine the validity of each separate item. Each item is analyzed with respect to an external criterion of job success, and the correlation of the item with the criterion is determined. Assuming that we now have a validity coefficient for each item, how shall we use that information to best advantage?



As soon as the validity is determined for each item, we are in effect treating each item as a separate test. It is as if we had validity coefficients for a very large number of very brief tests. Our problem is to combine those very brief tests in such a way that the composite will have the highest possible correlation with our criterion. The factors which make an item desirable for inclusion in a test are the same as those which make a test desirable for inclusion in a test battery, that is, validity and uniqueness. An item (or a test) will be esteemed as a valuable addition to the other items (or tests) in proportion to (1) its high validity and (2) its low correlation with the other items (or tests). The virtue of high validity in the single items is directly apparent. The importance of low intercorrelations is equally great, though less obvious. An item will add to the validity of an existing pool if it covers aspects of individual performance *not already covered* by that pool of items. This uniqueness is represented statistically by low correlation with other items. The effect of intercorrelation as applied to tests, which is exactly the same as the effect applied to items, was elaborated in Chapter 7.

If item intercorrelations were available as well as item validities, it would theoretically be possible to apply to items the same technique of multiple correlation and regression weights which was outlined in Chapter 7 for determining the optimum weights for combining full-length tests in a battery of tests. A regression weight could be determined for each item, to indicate just how heavily it should be weighted in arriving at the total test score. Several factors make this impossible in practice, however, and raise questions as to its theoretical desirability. Analytical treatment of item validities by the procedures of multiple regression must be ruled out for the following reasons.

1. The large number of items in a test usually makes it impractical to get the complete table of item intercorrelations. With only 100 items in an experimental test form, the labor of computing 4950 correlations of each item with every other would be an almost impossible task and one in which the labor involved seems out of all proportion to the gain.

2. Even if the intercorrelations were all available, the remaining labor of determining the regression weights from the matrix



of intercorrelations and validity coefficients would be quite overpowering. It would be an extraordinary computer indeed who could view with equanimity the prospect of carrying out a Doolittle solution of a 100-variable problem, or even of solving such a problem by approximation procedures.

3. In the usual test blank, refined weighting of items is not possible. In most cases, the only point at issue is whether to include an item or not (i.e., whether to weight the item 1 or 0). Practical convenience in test administration and scoring argues against differential weighting of items, and evidence for the value of such refined weighting has not been sufficiently convincing in most cases to overcome considerations of practical convenience.

4. The characteristics of item validities are such that item weights would be quite unstable. If the items in a test blank are examined, they will be found to cover a rather narrow range in validity coefficients. An item with a validity coefficient as high as 0.25 or 0.30 usually represents an outstandingly valid item. The whole range of item validities from the most to the least valid may cover no more than 20 or 30 points. Within this narrow range, sampling fluctuations play a large part in determining which items happen to have highest validity in our particular sample. These sampling fluctuations are magnified in the determination of regression weights. With the resulting degree of instability fine discriminations of item weights seem hardly to be justified.

If we abandon the complete analysis of item validities and item intercorrelations as an impractical procedure for determining item weights, we are thrown back on some approximation procedure which will be simpler but still reasonably accurate. Our problem is to select from a pool of  $n$  items some group of  $k$  items, where  $k \leq n$ , which will, when each item is given the same weight, give a score which has the maximum correlation with the criterion. We shall consider three possible ways of proceeding.

The simplest approach is to ignore item intercorrelations entirely. The choice of the items to be weighted is then based exclusively on the validity of the single items. Some standard of validity is chosen. Any item is weighted if its validity is as

great as or greater than the specified standard. If items are found that have validity coefficients numerically as large as the standard but with negative sign, those items may be given negative weight if negative weights are permitted in the test under consideration. For some types of tests, such as biographical data questionnaires, it will be appropriate to analyze each response option for each item separately, and to key any response option which has a validity coefficient whose absolute size without regard to sign comes up to the specified minimum. Take, for example, the following item:

At what age did you first go out with girls?

- A. 14 years old or earlier.
- B. 15 or 16 years old.
- C. 17 or 18 years old.
- D. 19 years old or older.
- E. Have never gone out with girls.

It might be found that options A and B had positive validities which came up to the specified standard, option C had substantially zero validity, and options D and E had negative validity up to the standard. The scoring of this item would give a plus credit to individuals who chose either response A or B, no credit on response C, and a negative credit for responses D or E.

If item intercorrelations are ignored, the only problem to be faced in item selection is that of setting the standard for including items. The problem is to select a level of item validity which will yield the maximum validity in the composite score. The addition of certain items with unit weights may add nothing to the validity of the test score, even though the validity of the items is positive. This may be illustrated by a hypothetical example. Suppose we have 5 items with validities of 0.25, 0.20, 0.15, 0.10, and 0.05, respectively. Let us see what validity we would obtain from a test composed of 1, 2, 3, 4, and finally all 5 of these items when the correlation of each item with each other is assumed to be first 0.50, then 0.20, and finally 0.00. Using the standard formula for the correlation of sums, we arrive at the figures shown in the table. An examination of this table

VALIDITY OF COMPOSITE SCORE FROM FIVE ITEMS OF DIFFERENT VALIDITIES

<i>Items Included</i>	<i>Size of Item Intercorrelations</i>		
	<i>.50</i>	<i>.20</i>	<i>.00</i>
1	.250	.250	.250
1 & 2	.260	.290	.318
1, 2, & 3	.245	.293	.346
1, 2, 3, & 4	.222	.277	.350
1, 2, 3, 4, & 5	.194	.250	.335

shows that the maximum validity for the total score is obtained when 2, 3, and 4 items, respectively, are included in the test score for the three illustrations chosen. In each case, inclusion of the least valid item or items results in a lower validity for the total score. Inclusion of the least valid items adds an amount of non-valid variance which outweighs any increase in the valid variance resulting from their inclusion in the test.

The problem is to select the group of items that will yield a score having maximum validity as applied to a new sample of cases. This involves, in addition to the problem of determining the critical value of item validity which gives maximum validity in the specific sample studied, a consideration of the regression of item validity coefficients from that sample to new samples.

Since the item intercorrelations are unknown and have an undetermined effect on the validity of the composite score, no analytical solution of this problem is possible. A sequence of operations for selecting a group of items to be retained and weighted is outlined in the following paragraphs.

1. When the validity coefficients for the separate items or item responses have been computed, the validities should be examined to determine whether the complete set might reasonably have arisen by chance. Because of the unknown correlation between items and especially between item responses, a rigorous test of this point probably will not be possible. However, if the distribution of item validities corresponds closely to what might be expected as a result of sampling fluctuations from a population in which all the validities were zero, any further detailed analysis devoted to selecting certain items for weighting is almost certain to be futile.

2. Given that the distribution of validities seems not to be attributable to chance, several scoring keys can be set up, setting different levels of validity as the minimum for an item to be weighted. Thus, four keys might be tried, the first weighting all items with a validity of  $\pm 0.15$  or over, the second all with a validity of  $\pm 0.12$  or over, the third all with validity of  $\pm 0.10$  or over, and the fourth all with validity of  $\pm 0.08$  or over.

3. The papers can then be scored with each of the keys described above. The score resulting from each of the keys can be correlated with the criterion, and it can be determined empirically which key gives the score with the highest validity.

4. In choosing the key to be recommended for use on new samples, consideration must be given to the regression which will occur as we move to a new group of individuals. If the scoring key is composed almost exclusively of items whose validity is of the same sign (all plus or all minus), it is desirable to use a key that sets a standard of validity somewhat lower than that which yields maximum test validity for the specific sample analyzed. The tendency of items to regress toward the group average makes the items with highest validity relatively less outstanding in the new sample, and maximum validity can be achieved by weighting a somewhat larger group of items. However, if the scoring key includes both positively and negatively weighted items in somewhat comparable numbers, the standard of validity which gives the key with maximum validity for the sample studied may be expected to give the key which will also be most valid for new samples. Regression of item validities takes place in this case also, but since the average of all item validities is approximately zero and regression takes place toward that value, the proportional size of the validities for different groups of items will be maintained.

The second general attack upon the problem of selecting items to be weighted in a scoring key takes account of item inter-correlations through a series of approximations. The general procedure is to set up an initial key for the test by keying a rather small pool of the most valid items. A score is obtained for each individual, based on this key. The validity of these scores is determined. Then the correlation of each item with this score is obtained. That is, the score on the new key is used

as the continuous variable for the item analysis.<sup>6</sup> We then have for each item (1) its correlation with the criterion and (2) its correlation with the initial pool of weighted items. We also know the validity of the score based on the initial pool of items. Given these values, we can compute for each item its *partial correlation* with the criterion when score on the initially chosen pool of items is held constant. This tells us what additional contribution each item would make to the validity of the initial pool. The analysis for each of the items already in the pool tells how much that item would add to the rest of the items in the pool.

We can now choose a second pool of items to be scored, selecting those with the largest partial correlations. This selection takes account not only of the validity of the single items, but also of their correlation with the initial pool of scored items. It consequently comes closer to being the best set of items. The whole process can now be repeated, if desired, scoring the tests on the second pool of items, correlating each item with this second pool score, and determining the contribution which each item makes to this second pool. It has been reported, on the basis of a procedure resembling this one, that the pool of items stabilizes quite promptly, and that further approximations after the first or second contribute very little to the validity of the resulting key.<sup>7</sup>

In this procedure, the problem of how rigorous a standard to set for item inclusion remains. If the test is such that only positive weights are to be considered, a standard procedure would be to include at each approximation all items whose

<sup>6</sup> For items that were included in the original pool, the correlation should be corrected for the spurious element introduced by including the item itself in the composite. This is accomplished by the formula

$$r = \frac{r_{it}\sigma_t - \sigma_i}{\sqrt{\sigma_i^2 + \sigma_t^2 - 2\sigma_i r_{it}\sigma_t}}$$

where  $r_{it}$  = the correlation of the item and the total score.

$\sigma_t$  = the standard deviation of the total score.

$\sigma_i$  = the standard deviation of the individual item.

<sup>7</sup> This method is a variation of one proposed by Flanagan. See J. C. Flanagan, "A Short Method for Selecting the Best Combination of Test Items for a Particular Purpose," *Psychol. Bull.*, **33**, 603-4 (1936).



partial correlations with the criterion were positive, when correlation with the previous key was held constant, no matter how small they were. This appears to be an unduly liberal procedure, however, and probably would not yield a final key with the maximum validity. If both positively and negatively weighted items are to be included, the sign of the partial correlation cannot serve as a guide to item inclusion, and the absolute magnitude of the partial must be used instead. The problem becomes one of determining at what size of partial correlation the additional valid variance covered by the item balances the additional non-valid variance which is introduced by adding the item with unit weight. As in the previous method, several keys may be prepared, each based on a different standard of size of partial correlation, and the resulting validities may be compared empirically. This is a laborious undertaking but is probably the only defensible one unless some rational basis can be established for the choice of cutting value.

A third alternate procedure for approximating the best set of items starts with the total set of available items, rather than a small pool of the most valid, and prunes off the useless ones.<sup>8</sup> Initially, the validity of each item is determined. A scoring key is prepared for the complete test. For tests of knowledge or ability, this key may be prepared a priori. For measures of interest, adjustment, personal history, and the like, it is prepared on the basis of the item validities. The key may be limited to positive weights, and probably would be in tests of ability, or it may include both positive and negative weighting of items. Once the key has been prepared, each test paper is scored with it. Correlations of each item with total score on the test are now obtained. The correlation of total score with criterion is also required.

Given the above data, we may proceed to calculate the partial correlation of each item with the criterion when score on the total test is held constant. (The total test usually includes a large enough number of items so that the inclusion of each item in the total score with which it is being correlated will hardly have an

<sup>8</sup> This approach is a variant of one proposed by Horst. See A. P. Horst, "Item Selection by Means of a Maximizing Function," *Psychometrika*, 1, 229-244 (1936).

appreciable effect.) A reduced test may then be prepared by eliminating those items with the least promising partial correlations. If negative weights are being used, these will be the items for which the *absolute* size of the partial correlations is least. If only positive weights are being used, all items with negative partial correlations will be the first to be eliminated.

The elimination of items may well proceed in two or even more stages. That is, as a first approximation the most unpromising items are eliminated. The reduced test is then scored for each individual. The validity of the reduced score is determined, and the correlation of each item with the reduced score is computed. Partial correlations with the criterion, holding total score on the reduced test constant, provide the basis for further pruning out of the least effective items. This stepwise procedure is more exact than a single elimination of items because the total score used for the second pruning will correspond more closely to the score on the final refined test. The correlations on the basis of which certain items are rejected will correspond more closely to the correlations with the final effective test.

Once again, we face the problem of how many items to screen out of the original item pool. In some cases the decision may be dictated by the desired length of the final test. Administrative considerations may dictate the inclusion of the best 75, or 100, or 150 items. The partial correlations in this and the previous method then serve to identify those items. When no such consideration exists, a feasible procedure is to decrease the size of the total pool of items by successive steps. With each successive pruning, the correlation of total score with criterion score may be determined. When this correlation commences to decrease, it is time to stop further reductions in the pool of items. Because of regression effects, we may anticipate that, for continued use in new samples, maximum validity will be obtained from a pool of items somewhat larger than that yielding maximum validity in the specific sample analyzed.

The procedures just discussed for item selection on the basis of item validities represent rather substantial undertakings, in terms of the amount of computation which is involved. They can probably be justified only when the item validities are based on a large group and when it is anticipated that the test items

will be extensively used, so that the investment of a good deal of time is justified to arrive at the best selection from among them. These procedures have probably never been given an adequate empirical trial, so that it is not possible to estimate how much improvement in validity the more refined procedures may be expected to add to rough, intuitive procedures of item selection.

### USE OF INTERNAL CONSISTENCY DATA

In tests made up of a more or less homogeneous group of items, in which each item purports to tap the same psychological functions as the others, it is appropriate to examine the internal consistency of the individual items. We determine the correlation of each item with the total score based on the whole set of items. This type of analysis, which can be carried out as soon as a preliminary form of the test has been administered to an experimental group, provides information of value both for the selection of test items and for the editorial revision of items. We must now consider the use of internal consistency data in these two ways.

The logic of using internal consistency data for item selection is somewhat more complex than the logic of using item validities for this same purpose. The internal consistency of single items is related to the homogeneity of the total test. The higher the internal consistencies are, the more homogeneous will be the test. But how high a degree of homogeneity is desirable in a test? In the case of validity, this question does not arise, since validity is the ultimate goal in test construction. The more valid the total score in the test, the better satisfied we will be. However, homogeneity in a test is not an end in itself. It is a means to attaining validity in the test score, to attaining validity in a battery of tests, or to attaining a degree of analytical clarity in the interpretation of the score on the test.

An item that shows a very low correlation with total score on a test is either very unreliable or measures functions quite different from those measured by the rest of the items in the test. In either case, items showing very low correlations with the test as a whole, and particularly those showing negative correlations,

are probably undesirable items. They should either be rejected completely or else revised and tried out again. If we are interested in developing a test which will yield a single score with some unity of meaning, a score to which a single name can reasonably be applied, the items at the low end of the scale for internal consistency are clearly the least desirable. Is the reverse true? Are items with highest internal consistency generally the most desirable?

The item showing high internal consistency *must* have a substantial degree of reliability. This will be an attractive feature of such an item. It must also have a large amount of overlapping of the functions covered by the other items in the test. Up to a point this is good, since the test is supposed to have a substantial amount of unity and coherence. However, a test is also supposed to have a certain amount of breadth and scope, and should cover in a representative fashion the scope with which it deals. Exclusive preoccupation with item internal consistency may lead to an undue narrowing of the scope of the test.

Suppose that a preliminary form has been prepared for a survey test of knowledge of physical science. Let us assume that of the 200 items in this test 80 deal with physics, 80 deal with chemistry, 20 deal with geology, and 20 deal with astronomy. Given any specialization of knowledge, we may expect greater correlation between items within the same field than from one field to another. Because of their predominance in the total score, we may expect the physics and chemistry items to show higher item-test correlations. It is entirely possible that if we select from the total pool the 100 items with highest internal consistency we shall limit the final test exclusively to physics and chemistry items. This represents a narrowing of range of content which may be quite inappropriate to the original purpose of the test. Internal consistency data must be used with discretion within the framework of the original outline and specifications for a test. They cannot override the outline, and they do not provide a substitute for a definite content outline for the test.

The selection of items on the basis of internal consistency contributes to the reliability and the homogeneity of the resulting test score. Empirical evidence supports our previous discussion

in indicating that this information is most effective in the negative sense. That is, internal consistency data are more useful as a means for the discovery and elimination of the usually rather small percentage of unsatisfactory items than as a means of selecting a small fraction of items that can be identified as *the best*. Empirical studies show that tests made up of items with intermediate internal consistency values have very nearly the same reliability as tests made up of items with the highest internal consistency. It is only when items in the lower ranges of internal consistency are used that reliability of the resulting test suffers.

Our discussion in previous paragraphs suggests that there may sometimes be a question as to what should be used as the total score, against which the single items should be analyzed. In the illustration of the physical science survey test, we might inquire whether we should be interested in the correlation between an astronomy item and score on the total of all the items of all types or the total of all astronomy items. The problem is that of deciding how broad a range of content and function should really be thought of as homogeneous. Logically, we are usually on stronger ground when the total score represents a single narrowly defined type of test material. Items may then be correlated with subtest totals, where the subtests are composed of a single item type. Homogeneity is a matter of degree, after all, and there is no entirely clear-cut definition of what constitutes a single unified test.

Internal consistency data should be used in item selection first of all to screen out the definitely unsatisfactory test items. Beyond that point, the data should be used, in connection with item difficulty data and with the outline of specifications for the test, to select the more promising items in each specified segment of content or function measured. In addition, internal consistency item analysis data provide a most valuable source of information to the item writer in revising and improving the items which he has written. For this purpose, data are needed with regard to each of the response options on an item. The analysis should show the percentage choosing each option separately for the upper and lower fraction (usually 27 per cent) of



the group on the basis of total score. These data may be used in revising and rewriting the item. Such rewriting and revision may be worth while not merely to salvage poor items but also to improve details of items which are in general satisfactory.

Consider the following item for a test of chemical information:

If copper and zinc are melted down together and the molten mixture is allowed to cool, the product will be

- A. An alloy.
- B. An amalgam.
- C. A new chemical compound.
- D. A salt.
- E. Monel metal.

Suppose that analysis of the upper and lower 27 per cent of a group completing a course in high-school chemistry gives the following results:

Choice	Per Cent Choosing	
	Top 27%	Bottom 27%
Option A	79	53
Option B	12	8
Option C	5	18
Option D	0	1
Option E	3	15
Omitted	1	5

A study of these results shows that option D was attracting almost no choices from anyone in either group. This suggests that option D may well be replaced by a more attractive mislead, if one can be thought of. The test editor would rack his brain for a more plausible wrong answer to substitute for D. Option B is attractive, but it attracts the wrong people. It draws more choices from the upper than from the lower group. It looks as though the word "amalgam" is meaningless to many in the lower group, whereas those in the upper group have enough of an idea of its meaning to confuse it with "alloy." By its nearness in meaning to the correct answer, this mislead lowers the discrimination of the item in the group studied. The test editor would consider finding a different option to take the place of option B.

In general, a study of the percentages choosing each response option will reveal two types of inadequacies, as illustrated in the preceding example. Some response options are non-functioning, i.e., they are not attracting any choices from either group. Some misleads fail to discriminate or discriminate in the reverse direction, i.e., they are chosen as often or more often by the upper as by the lower group. Each of these provides a cue to the test editor for revision of the item. This revision often takes the form of replacing the deficient misleads. It may also sometimes result in rewording of the stem of the test item.

## *The Administration of a Testing Program*

Any personnel selection program involves administrative as well as technical problems. The administrative problems bulk largest in the continuous routine of testing operations from week to week and month to month. However, they also arise in the collection and analysis of research data. Making plans for collecting criterion data, arranging test records in such a way that they can conveniently be collated with criterion data obtained at some later date, scheduling research testing so as to interfere as little as possible with other activities and so as to elicit the best rapport of subjects all represent administrative aspects of research testing. Problems of scheduling, flow of subjects, procuring, security of test materials, scoring and auditing of tests, and reporting of results represent administrative problems of routine testing.

Though administrative problems arise in any testing program, they become progressively more important as the program becomes larger. When all the testing and all the subsequent activities with the test results are being carried on by two or three persons, all located at one point and all working closely together, work can proceed relatively informally with a minimum of formalized planning. But in a large-scale program, of the type represented by the work of such agencies as the College Entrance Examination Board or by the testing of military personnel during World War II, procedures must be definitely specified in black and white and planned in advance in much more detail. Where testing is to be carried out in a number of places and by large numbers of individuals who are not specialists in the work, the only hope for uniformity of procedure lies in careful administrative organization and a detailed manual of operating procedures.

Effective administrative organization has two major goals. The first of these has already been mentioned. It is the achievement of uniformity of procedure in test administration, scoring, and weighting, so that the final evaluation of a given individual will be the same no matter where, when, or by whom he was tested. The other goal is that of efficiency of operation.

One evidence of this efficiency is the promptness with which the test results become available for routine use. The pressure for prompt availability of test results varies from one instance to another. A college, testing applicants for admission, may have several weeks in which to reach a decision about each individual, whereas an employment office may have to give an answer the same day. In spite of such variations, however, it is quite usual for a testing program to operate under considerable pressure for prompt availability of the results, so that an administrative decision may be reached with regard to each man. When several hundred men are being tested each day with a complex battery of tests, an efficient and smooth-running organization is required to turn out accurate composite scores for predicting job success within a day or two of the time each man was tested.

Efficiency shows up, in the second place, in the readiness with which the test results may be used for reference and research work. As a testing program carries on over the months and years, test records accumulate for a large number of cases. Though civilian programs will hardly approach in size the seven or eight million who took the Army General Classification Test or the more than half a million who took the battery of twenty Army Air Forces air-crew classification tests during the war, the numbers tested in any sizable industrial, civil service, or educational testing program may well run into many thousands. If a continuing program of research is to be carried out with these tests, and if they are to be validated against later criteria of job success with a view to improving the selection procedure, the results must be in convenient form, efficiently organized, and readily accessible for use.

Finally, efficiency is evidenced by economy. An efficient organization accomplishes a given program of testing with the least expense and keeps waste to a minimum.

In this chapter we shall consider some of the administrative problems which are likely to be encountered in any extensive testing and research program. We shall consider first administrative problems in the conduct of testing. Then our attention will be turned to the problems of scoring tests and combining test scores into composite scores for predicting job success. Finally, we shall examine some of the factors encountered in setting up an effective system of reports and records. Specific problems will be outlined, and some suggestions will be given for dealing with these problems.

In the discussion which follows, it is assumed that a fairly extensive battery of tests is being given to substantial groups of subjects. Some of the points obviously will not apply if only one or two tests are to be given, or if testing is being carried out with only one or a few individuals at a time.

## ADMINISTRATIVE PROBLEMS IN THE CONDUCT OF TESTING

### *Schedule of testing*

One of the first things to be decided in any testing program is: When and for how long is each applicant to be tested? The answer to these questions, particularly the question "How long?," sets the framework for the whole testing program. It delimits the final battery which may be used and may determine to a considerable extent how much of a research program of test development and validation can be carried on. The personnel psychologist will usually have some part in the decision concerning the amount of time to be made available for testing. He may well be asked how much testing time he needs to obtain satisfactory data for the desired selection or classification, or he may be asked to submit a proposed program of testing. However, the final decision is usually made at a level of management or command one or several steps above the psychologist. The psychologist proposes, but top management disposes. The final time allowance is likely to represent a compromise between the interests of the over-all management, which wants a simple, inexpensive, but accurate method of personnel classification, and



the psychologist, who knows that accurate determination of aptitude for one or several jobs is not a simple matter. The actual time available will be a function of the aggressiveness and persuasiveness of the psychologist, the prestige of personnel testing in the eyes of management, and the importance of the job for which selection is being carried out.

In any event, the program of testing will function within certain over-all limits of time to be made available for testing. The psychologist usually has no difficulty in persuading management to use a shorter test battery, but he ordinarily has great difficulty in getting authorization for a longer one. The testing time needed to achieve somewhere near the maximum effectiveness of prediction for one job or for a group of jobs is almost impossible to estimate in advance. A sound judgment on this point is one of the end products of a program of research, rather than one of the initial premises of a testing program. It is perhaps sound policy, therefore, to request as large an allowance of time for testing as the traffic will bear in a particular practical situation.

Testing time will be limited on the one hand by the amount of time that can reasonably be made available for each subject and on the other hand by the facilities available for handling the test results. There is, of course, no point in giving more tests than can be scored and used either as a basis for evaluating the individual or for further research on prediction of that job. If facilities are available for analyzing the results, there will usually be enough tests meriting trial on a research basis, either existing standardized tests or research tests built especially for the testing program, to fill any amount of available testing time.

The particular time for testing is usually fixed by practical administrative conditions. Most of the considerations in setting testing times are those practical common-sense ones which any sensible person is likely to anticipate. Except for very particular and unusual purposes, individuals should be tested when they are fresh, rested, and feeling their best. Although this point may not be so important for actual test scores as anticipated, it is important for the general morale of the testee both during and after the testing. It is important for good will toward the testing program that each person tested feels that he was fairly tested.

It is advantageous that the testing periods not be too long. Subjects should have an opportunity to relax and to get relief after an hour or two of testing. The periods should not conflict with normal meal times. On the other hand, testing periods should not be too broken up, since this wastes time for all concerned and may be somewhat disturbing.

In a continuing testing program, it is often desirable to test an individual as soon after his original contact with the organization as possible. This minimizes the individual's chances to hear about the tests from others who have been tested and thus makes for a truer comparison of different individuals' abilities. Where testing occurs only at intervals, the general schedule into which the tests and test results must fit will determine the dates for testing.

The sequence in which the separate tests in a test battery are given is usually governed by factors of convenience. It is usually convenient to group together all tests requiring the same sort of facilities. Thus, if there were some printed group tests and some individual apparatus tests, one would usually give all the group tests before starting the individual tests. Similarly, motion picture tests requiring darkening of the room and projection facilities would usually be grouped as a separate unit. Further minor factors of convenience will probably determine the detailed sequence. Tests should be arranged so that a break for a rest or for a meal comes conveniently between two tests. Short speeded tests may often be alternated with longer tests with more of a power element, both to provide variety for the subjects and to facilitate distribution and collection of materials. A sequence to provide variety in the content and type of activity in different tests may also be desirable, from the viewpoint of maintaining interest and effort by the subjects.

### *Administration of printed group tests*

Printed group tests are undoubtedly the most economical and the simplest to administer in an objective and professional manner. They make few demands either on facilities or on testing personnel. The primary material requirement is a satisfactory testing room. The primary personnel requirement is a well-trained test administrator who is adequately supported by

proctors. Other administrative provisions concern the organization of material and supplies and the instructions to the subjects.

The ideal room for the administration of group tests has the following characteristics:

1. It is quiet and free from the disturbances of other activities.
2. It is well lighted and ventilated.
3. It provides each subject with a comfortable seat and a good writing space, preferably a desk or table. A chair with a writing arm is reasonably satisfactory.
4. It has appropriate size and shape and has sufficiently good acoustics, so that each person being tested can both see and hear the test administrator without difficulty.
5. It provides space so that test proctors can reach any subject being tested, to answer a question or inspect his work.
6. It provides enough separation between testees to make cheating difficult or impossible.

7. Where testing is of large groups and for considerable periods, the testing room has adequate nearby toilet facilities.

Testing will often have to be done, and can be done quite successfully, in a room that does not come up to these specifications. However, they provide a standard towards which to work.

Of the testing personnel, the most important is the administrator in charge of testing. For him, the important attributes are "presence" in front of the group, a thorough familiarity with the instructions which he is giving, conscientiousness in following out the specified procedures with exactness, and judgment in the handling of the occasional incidents for which no rules are provided. "Presence" includes assurance and poise before a group, a degree of dominance which enables him to remain in control of the group, and a good speaking voice. Concerning instructions more will be said presently. The administrator should be well acquainted with the verbatim instructions for each test and with the rules of testing procedure. He should be able to give the instructions accurately without continuously referring to them, and he should know the rules for answering questions on timing tests with meticulous care. Finally, he should have sufficient understanding of the general principles and purposes of the testing so that he will be able to exercise good judgment

when faced with any situation which is not explicitly covered in the testing procedures.

When testing is carried out with large groups, the test administrator requires the assistance of test proctors. The number of these depends on the size of the group, the complexity of the testing, and the conditions under which testing must be carried out. More assistance will be needed for a program involving a number of tests, with much passing and collecting of test blanks and frequent instructions for new test tasks. More proctoring will be needed in a room in which the physical arrangements make cheating easy. In general, the functions of proctors are the following:

1. To pass out and collect all necessary test materials.
2. To answer individual questions on procedure, in accordance with and within the limits of the standard testing procedures.
3. To check the subjects, to make sure that they turn pages when they are told, stop where and when they are told, are working according to the instructions and on the appropriate part of the test.
4. To discourage cheating by their presence and to stop it if it occurs.

The test instructions, both those which are given to the subjects and those which guide the test administrator, are among the most important aspects of test procedure. For power tests of a type familiar to those being tested, instructions and procedural details are perhaps not of great importance. Almost any public school graduate knows at once how to respond to the common types of test of word knowledge, reading comprehension, or subject matter information, and, even if he does not know immediately, if time is ample he can be expected to figure it out without suffering a penalty in score. When the test is of a more novel type, however, and when it emphasizes speed of performance, the content and presentation of the instructions become increasingly important. It is important then that the instructions be full and clear and that they be uniformly and consistently presented. In novel types of test material, score may depend to a great extent on understanding what one is supposed to do. Unless the test is expressly designed to test comprehension of instructions, the instructions should be of such

a nature and so presented that everyone has the best opportunity and the same opportunity to understand them.

In preparing written instructions and supplementary statements of procedure to guide the test administrator, the following principles may well be followed:

1. Instructions should be full. They should err in the direction of boring repetition rather than the reverse. It is better for some testees to be bored for a moment or two than for others to miss some important point. Human beings have an almost unbelievable ability not to follow instructions, and even with repetition some individuals will almost certainly get mixed up on any novel or unusual procedure.

2. Instructions should be in simple language. The vocabulary should be familiar and sentences should be short. Comprehension of what is being said should present as slight an obstacle as possible to the examinee.

3. Except where the procedure is extremely simple, instructions should include illustrative examples and usually a practice exercise. These represent one of the most effective techniques for guaranteeing understanding of the test task, or for discovering and correcting misunderstanding if it is present. A practice exercise takes time away from the total time available for testing, but the time is well invested if it equates factors of understanding of the task and technique for attacking it.

4. The basic instructions should usually be printed on the test booklet. These often need to be supplemented by verbal instructions from the examiner. Such verbal instructions should be uniform for each group of subjects and should usually follow verbatim a typed manuscript.

5. In particular, illustrations, examples, and suggestions as to technique of responding to items should be kept scrupulously uniform from group to group. In some tests, particularly highly speeded ones, a hint as to the most effective procedure for attacking the problems may make a very real contribution to score. Such a hint should be always given or always omitted. It should never be left up to the whim of the examiner.

6. On any but the simplest types of test, subjects will raise questions. A standard procedure should be adopted for dealing with these. In a continuing testing program, the more common



questions are soon identified, and a standard procedure for answering them can be written up and learned by each person in the testing room. These standard answers should be supplemented by a statement of principles to guide the answering of the less common questions or for dealing with new tests or research tests.

The preparation of materials for use in group testing is one administrative detail which cannot be slighted if a mass testing program is to proceed smoothly. With reusable booklets, this preparation starts as soon as booklets are returned from the testing room. The count of booklets must be checked, to make sure that all have been returned. Booklets must be scanned, and any marks removed from them. Any booklets which have marks that cannot be removed, or which have become too worn or soiled, should be withdrawn and destroyed. It is often desirable to slip the answer sheet into the test booklet, so both can be passed out at once. Where special pencils are being used, as is required for machine scoring, these should be checked to make sure that they are operating properly and have a sufficient supply of lead. Sufficient quantities of all materials for the coming day's testing should be counted out and arranged in a storage room convenient to the testing room.

### ***Apparatus test problems***

For the measurement of some aptitudes and skills it may be necessary to resort to individual tests, using apparatus of some sort rather than a printed test blank. The apparatus may be no more than a simple pegboard, or it may be a much more complex mechanical or electronic contrivance. The apparatus will be used to assess some aspect of behavior which seems untestable with a printed test. The personnel psychologist may well be somewhat reluctant to introduce apparatus tests into his battery. There is first of all the matter of economy. A test that can be given to only one subject at a time by an examiner, or at best to four or five, is a much more expensive item to include in any testing program than a test that can be given to groups of a hundred or two hundred subjects by a team of three or four men. In addition, with the apparatus test the problem

of maintaining uniformity of testing conditions for each man tested becomes much more serious.

Uniformity of conditions for each man tested with an apparatus test has two aspects. The first of these is uniformity of procedure by the examiners, and the second is uniformity in the apparatus itself. Uniformity of examiner procedure is more difficult in apparatus testing than in group testing because (1) the apparatus test is likely to rely more on the human examiner and less on printed material to give instructions to the examinee, (2) instructions are given anew for each subject tested, and (3) almost of necessity a larger group of individuals serve as examiners, making for more personal differences between examiners. It becomes necessary, therefore, to put more emphasis upon the training of examiners in the administration of individual tests. The procedure for giving instructions should be learned verbatim by each examiner. Procedure should include not only what the examiner says but also what he does, i.e., the amount and type of demonstration which he gives. Procedures for scoring and recording of results should be made as mechanical as possible. At this point it should be noted that, since an apparatus test usually does not leave a permanent record of performance comparable to the answer sheet of a printed test, which can be checked or rescored at leisure, special emphasis should be placed on checks on the score at the time it is recorded.

Maintaining uniformity of the apparatus from person to person represents another problem in apparatus testing. This is especially true when several copies of an apparatus are used. Experience in the AAF air-crew testing program, which probably represents the most extensive use of apparatus tests in test history, showed very clearly that apparatus differences can be quite substantial even for well-constructed equipment. It was found necessary to prepare very detailed specifications and to build apparatus to high standards of precision if comparability was to be achieved from one piece of apparatus to another.

Even maintaining comparability of scores within the same piece of apparatus from day to day and week to week can represent a real problem. Loosening up of bearings, wearing of contact points, accumulation of dust and grit can change the performance characteristics of a piece of equipment and cause

a progressive shift in the distribution of scores. If a piece of apparatus is to function in a consistent way, it must be subjected to a rigorous program of preventive maintenance and of calibration. A routine of oiling bearings, cleaning electrical contact surfaces, checking motors, and the like must be worked out. Furthermore, the apparatus must be studied to see what contacts, pressures, etc., are critical in determining scores, and a program of calibration checks devised to see that these adjustments are kept within close limits.

In the AAF, even with systematic maintenance and calibration, it was also found necessary to keep continuous records of scores for successive groups of subjects. The mean score was computed for each one hundred subjects for each copy of an apparatus test. If the score for a particular copy diverged too much from the general average of all copies of the test, that copy was subjected to a particularly thorough inspection and servicing. If the copy continued to give aberrant scores, it was removed from service until it could be gone over by an apparatus specialist.

The problems in using apparatus tests discussed above are real, but they do not negate the value of the tests. Even a fairly substantial amount of apparatus and examiner variance will make only a moderate reduction in test validity. The apparatus test retains much of its value in spite of these variations. However, it is well worth while to hold the variation to a minimum.

### *Motion picture test problems*

The use of the motion picture as a medium for testing still represents rather a novel venture. However, for certain aspects of testing this medium presents very real advantages. The advantages and problems of motion picture testing have been considered to some extent in Chapter 3. Since motion picture testing is ordinarily group testing, the administrative aspects of testing are much the same as for printed tests. However, sound motion pictures present both certain advantages and certain problems from the administrative point of view. The chief advantage lies in the complete objectivity and uniformity with which both instructions and test can be given. Once the picture sequence and sound track have been recorded, the uniformity of phrasing,

of timing, and of demonstration are guaranteed for every testing. The variable human factor is reduced to a minimum.

The problems which arise have to do with seating and illumination. Seating involves the factors of angle and distance from the screen. As the viewer is moved away from directly in front of the screen to the side, the shape of the pattern which is presented to his eye becomes more and more distorted. As he is moved back from the screen the retinal image becomes smaller and smaller. Fortunately, a good deal of perceptual constancy seems to operate in viewing a motion picture screen, so that objects are seen as having their proper size and shape, in spite of distortions in the size or shape of the retinal image. However, the resolving power of the human eye is limited. In any test which depends on seeing fine detail, greater distance from the screen will penalize the subject. As a matter of reasonable precaution, the tester should investigate the effects of distance and angle upon score for any motion picture test which he proposes to use. Where these factors do make a difference, the tester may either make provision for keeping these factors within rather narrow limits or develop correction tables to allow for the effect of position within the test room. In general, motion picture tests put somewhat severer limitations on size, shape, and other features of the testing room than do printed tests.

Illumination, together with the corollary problem of ventilation, constitutes the other practical problem in motion picture testing. It must be possible to keep the room dark enough so that a good image is formed on the motion picture screen, and yet light enough so that each subject can see to mark his answer sheet. In practice, it has been found that over quite a range of illumination levels this is possible, particularly if the lights are shielded so that they do not shine toward the motion picture screen. Any room which is to be used for motion picture testing needs to be equipped with appropriate shades and ventilators so that air can be admitted without admitting light.

### *Security of tests and testing materials*

In any testing program in which applicants are strongly motivated to achieve high scores on the tests, the problem of maintaining the security of the test materials arises to some

degree. There is always the possibility of leakage, so that some of the individuals tested learn about the tests and about specific test questions in advance of being tested. The seriousness of this problem depends on a number of considerations. Perhaps most important is degree of motivation to excel on the test. In most Civil Service Commissions setting examinations to fill federal, state, or municipal jobs, the motivation is deemed sufficiently intense so that the same examination is never administered a second time. Somewhat the same situation holds for College Board, State Regents, and other academic qualifying examinations. This is an extreme step to prevent previous acquaintance with the examination and it introduces a number of complications. It requires the development of new test forms every time an examination is set. It almost precludes the experimental try-out of a test, with subsequent editorial revision based upon preliminary results. It complicates any continuing program of research and development. This procedure is feasible only for examinations that are given only at stated intervals and are then given simultaneously to large groups.

In the usual personnel testing situation, motivation is probably not so great as in civil service testing, and the probability of available organized help for the testee is certainly less. However, the problem of preventing leakage of information about the tests remains a very real one. The following suggestions for preventive measures are offered:

1. Applicants should be tested as soon as possible after the initial contact with the organization and before they have had a chance to meet and receive information from those who have already taken the tests.
2. Test materials should be kept under lock and key, except when they are being used, and only authorized individuals should be permitted to have access to them.
3. An accurate inventory should be kept of all test materials. Only as many booklets as are needed for testing on a given day should be brought out, and these should be checked before the examinees leave the room to make sure that all booklets have been accounted for.
4. Where separate answer sheets are used and a test booklet is used over and over again, each booklet should be given a



serial number. Booklets can then be arranged by serial number and checked readily as they are returned.

5. Booklets whose usefulness has ended should be destroyed, preferably by burning.

6. Subjects should be permitted to take no books or papers into or out of the testing room. All paper needed for calculations and the like should be supplied and subsequently collected.

7. If possible, several alternate forms of each test should be provided, and the forms should be shifted from time to time. This adds both to the labor of test construction and to the complexity of norming, scoring, and record-keeping procedures, but it goes a long way toward minimizing the possibility of specific advance preparation on the test materials.

### *Motivation of subjects*

When testing is being carried out on applicants for a position and satisfactory test performance is a condition of obtaining the position, obtaining adequate motivation is ordinarily not a problem. The very fact that the individual has applied for the position is usually a guarantee of sufficient motivation to give maximum performance on the test. There are two types of personnel testing situations, however, in which obtaining a satisfactory level of motivation may be a real problem. One of these arises when tests are being used for classification, the other when they are being used for research purposes.

When tests are being used for classification, several job categories are involved. These usually form a hierarchy of desirability. The order of desirability is individual to a certain extent and varies from person to person. However, there is often a good deal of consistency in the expressed preferences for the different job categories. Thus, applicants for air-crew training in the Army Air Forces almost universally desired to be pilots, with navigator generally chosen as second choice and bombardier third. Test results had to be used to assign many applicants to other than a chosen job. In such a situation, any tests which are readily identified as pertaining to the non-preferred category are likely to evoke an unsatisfactory level of effort on the part of some subjects.

This slackening of effort can hardly be avoided completely. There are two approaches to reducing it. On the one hand, the superficial resemblance of test materials to the non-preferred jobs can be kept to a minimum. In fact, an effort can be made to give every test a superficial appearance of relationship to the preferred task. This can be done, of course, only to the extent that it does not change the fundamental nature of the function being measured. However, changing the cover design and the phraseology of content in a test for an unpopular job specialty can reduce somewhat the obviousness of its purpose. The second approach is through the introductory statement prefacing the testing. This should emphasize the importance of doing well on *every* test. If there is a selection as well as a classification function to the testing, the selection function should be played up and the classification function played down. The effort should be made to have the examinees feel that any lapsing of effort is a risky thing and that their best interests are served by getting the best possible scores all along the line.

Research tests also stand in danger of eliciting only perfunctory response from some examinees. This is particularly true when they are given to individuals who have already been accepted for or who are already employed on the job. When research tests are given at the same time that selection or classification tests are given, there is ordinarily no need to make a distinction between the two, and the motivation of regular tests can be carried over to the research tests. Even if the research tests *are* identified to the examinees as such, the general unfamiliarity and seriousness of the testing situation is likely to maintain a satisfactory motivational level.

With individuals already accepted for the job, especially individuals who have been on the job for some time, the motivational situation is likely to be quite different. They are likely to see little personal advantage in taking tests. In fact they may resent testing and be suspicious of the motives which prompt it. If their position is such that they can do so with impunity, they may refuse to take the tests or take them in such a perfunctory way as to yield a meaningless score. Motivation in these groups is in part a matter of morale—the general morale of the organiza-

tion and also the specific reputation of the personnel psychologists for integrity and fair dealing. In part it is a question of convincing the examinees that they do have a direct stake in the results of the testing. This stake can be partly in general terms, in the improved efficiency of the total organization and resulting benefits to the individual members. The reality of such benefits to individual workers will probably appear rather slight in many industrial organizations. The stake can be partly in improved chances for individual promotion and special consideration. In introducing the testing project to the examinees, the examiner should both reassure as to any negative action and hold forth promise of positive action as a result of the testing. It need hardly be added that he should live up to any commitments which he makes, in so far as it is at all possible for him to do so. A statement which implies a good deal but guarantees nothing, and which may be quite generally useful, is that the scores on the tests will "be made a part of each individual's permanent record."

## ADMINISTRATIVE PROCEDURES IN TEST SCORING AND WEIGHTING

### *Procedures for expediting scoring*

After tests have been administered, the next major administrative chore is getting them scored. Where testing is carried out on a large scale, this becomes a very substantial part of the labor in the whole testing operation. Scoring as we are discussing it now refers to group tests yielding some form of written record of the answers. Most individual apparatus or performance tests are necessarily scored as they are given, by observing the amount of work done, the time required to complete the task, the proportion of time on the target, or some other index of behavior. Accurate scoring of such tests will be facilitated by having the record a mechanical one, if possible, so that the tester needs only to read a time clock, an electric counter, or some other instrument.

The first decision on the scoring of answer sheets is whether they are to be scored by hand or by machine. The chief con-

tender for machine scoring is the International Business Machines Corporation Test Scoring Machine. This machine can be obtained only on a rental basis and involves an appreciable financial outlay.<sup>1</sup> However, an experienced operator can score several hundred papers an hour on the machine, a rate far in excess of any procedure for hand scoring. Whether it will pay to plan for machine scoring depends primarily on the size of the testing load. If the machine would be used only a few minutes a day, it would not represent a sound investment; if it would be used the major part of each working day, it would certainly prove an economy. Somewhere between these two extremes, the balance of economy shifts from hand to machine scoring. Just where it shifts depends on a number of factors, some of which will be discussed in the section on machine scoring.

### *Efficient hand-scoring procedures*

If it appears that testing operations will be carried out on a scale sufficiently small so that hand scoring is the more economical procedure, attention should be turned to making hand scoring as efficient as possible. Much can be done to improve the efficiency and accuracy of the hand-scoring procedures often found in practice.

Hand scoring can be improved particularly by providing the proper printed form for use in recording answers and as a scoring key. If efficiency of scoring is a consideration, answers should always be recorded on a separate answer sheet. This may complicate slightly the task of the examinees, but the elimination of continuous turning of pages on the part of the scorer is very ample recompense. In addition, the provision of a separate answer sheet makes it possible to use test booklets over and over again. This economy will become a very substantial one when illustrations, color, or other characteristics of a test make copies expensive to buy or print.

The most efficient hand-scoring answer sheet will have every separate choice of a response alternative represented by a different *place* on the answer sheet. A part of such an answer sheet is

<sup>1</sup> Monthly rentals in 1948 were reported to be \$40 or \$60, depending on the type of equipment furnished.

represented in Figure 1. The subject then indicates his choice of answers merely by making a mark in the spaces corresponding to the answers which he chooses. The great advantage of this type of answer sheet is that it reduces scoring the paper to a simple operation of counting marks. For scoring purposes, a stencil is prepared for the right answers (and possibly one for the wrong answers). The stencil is made to fit the answer sheet and should be provided with conspicuous alignment marks, to

Course \_\_\_\_\_ Name \_\_\_\_\_  
Exam. \_\_\_\_\_ Date \_\_\_\_\_

Instructions: Read the directions on the test sheet carefully, and follow them exactly. For each test item, mark your choice for the correct answer by blocking out the letter which corresponds to the best answer for that test item.

Item	Answer	Item	Answer	Item	Answer	Item	Answer
1	A B C D E	26	A B C D E	51	A B C D E	76	A B C D E
2	A B C D E	27	A B C D E	52	A B C D E	77	A B C D E
3	A B C D E	28	A B C D E	53	A B C D E	78	A B C D E
4	A B C D E	29	A B C D E	54	A B C D E	79	A B C D E
5	A B C D E	30	A B C D E	55	A B C D E	80	A B C D E
6	A B C D E	31	A B C D E	56	A B C D E	81	A B C D E
7	A B C D E	32	A B C D E	57	A B C D E	82	A B C D E

FIGURE 1. Section of a typical answer sheet for hand scoring.

make sure that it is accurately placed with regard to the answer sheet.

At the location on the scoring stencil corresponding to each possible correct response on the answer sheet, a hole is punched. When only a single answer to a question is permitted, all answer sheets must first be checked for double marking. No credit is given for items that have a double mark. Then when the stencil is placed over the answer sheet, it is necessary merely to count the number of holes in which marks appear in order to get a count of the number of right answers. For those tests in which the scoring formula specifies a penalty for errors, it is also necessary to have a stencil with a hole corresponding to each possible wrong answer. Using this "wrongs" stencil provides a count of the total number of wrong answers. The specified fraction of this number can then be subtracted from the "rights" score to yield the final score.

All hand scoring of answer sheets should be checked. On each answer sheet, each scoring operation should be carried out in-



dependently a second time by a different scorer. Checking should include both the initial steps of counting right and wrong responses and the subsequent arithmetical computations which lead to the score. In cases of disagreement between the two scorings, the answer sheet will have to be scored a third time with special care, to determine which was the right score.

A number of special patented devices have been developed to expedite hand scoring still further. These usually require specially prepared answer sheets. One of the most familiar of these is the Clapp-Young self-marking answer sheet. This answer sheet is backed by a piece of carbon paper. On the back side of the answer sheet only the locations (square boxes) corresponding to the correct answers are printed. When a mark is made in an answer box on the front of the answer sheet, the pressure transfers a carbon mark to the corresponding location on the back of the answer sheet. When the carbon paper is torn off, it is a simple matter to count the number of marks on the back of the answer sheet which fall within boxes (right answers), and the number which do not fall within boxes (wrong answers). However, an answer sheet such as this requires a special printing job and is justified only for a test which is to be used in large numbers. Where a test is to be used in large numbers in many different places where scoring machines are not available, this type of answer sheet may be enough simpler than the separate scoring stencil to justify its use.

Once the most efficient forms have been worked out, scoring efficiency becomes a problem in selection, training, and supervision of scoring personnel. It must be recognized that scoring is a repetitive, routine, clerical task. It is conducive to boredom and requires attention to detail. Scoring personnel should be chosen with these points in mind. Relatively little training is required when the types of materials described above are used, but as the operation becomes less simple and mechanical some training should be provided in the required skills. Supervision means the scheduling of work assignments to different workers, maintenance of performance records for different workers, setting up a system of rest pauses, and the like. It may be remarked in passing that, if scoring involves as much time of as many persons

as this paragraph implies, it will probably be more efficient to carry it out with machines. These remarks are therefore perhaps more relevant to machine scoring.

### *Machine scoring procedures*

In the discussion of machine scoring, attention will be limited to the International Business Machines Corporation Test Scoring Machine because it is commercially available throughout the country and has received wide use. Other mechanical methods of test scoring have been devised from time to time, but none of these has reached general commercial distribution. In discussing the IBM Test Scoring Machine, space does not permit a full exposition of its principles of operation, the steps in scoring tests with it, or the operations which it will perform. This information can be obtained from the operating manual of the machine. We will attempt here to present a summary statement of (1) the principles on which the machine works, (2) its assets and liabilities, and (3) certain hints concerning its efficient use.

The IBM Test Scoring Machine is a device which uses the electrical conductivity of a graphite mark on a piece of paper as the basis for counting and summing the right and wrong answers to a test. It uses a special answer sheet with answer spaces located on the sheet in a fixed pattern. Part of an answer sheet is reproduced in Figure 2. The examinee fills in the spaces on the answer sheet corresponding to the items which he considers right, using a special electrographic pencil with a soft, electrically conductive lead. When the answer sheet is inserted in the scoring machine, a set of contact brushes is brought against the answer sheet. Each mark establishes contact across one set of brushes. The circuit of the machine is such that each contact causes a flow of one unit of electrical current. By using a scoring stencil, the current from the correctly marked items can be caused to flow through one circuit and the current from the wrongly marked items through another. These separate circuits have separate rheostats, so that the units for right and wrong answers can be varied independently of each other. The current from the "wrongs" circuit can be either added to or subtracted from that in the "rights" circuit. The combination of these last

two characteristics makes it possible either to add or to subtract any fraction of the "wrongs" score from the "rights" score. The final score is obtained by reading the position of the pointer on a galvanometer dial. Three independent circuits on the machine make it possible to obtain three separate part scores upon a single insertion of the answer sheet, or the total score may be obtained

NAME _____		DATE _____		DO NOT WRITE IN THIS SPACE
LAST	FIRST			
PLACE (SCHOOL) _____		CITY _____		
AGE	SEX	GRADE		
YRS.	BOY	BOY		
				RAW SCORE _____
				SCALE SCORE _____
				PERCENTILE _____
				NORMS USED _____

25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5
A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D
E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E

FIGURE 2. Section of a typical answer sheet for the IBM Test Scoring Machine.

directly. The machine has a number of other special features which cannot be elaborated here.

The steps in operating the scoring machine will be described only briefly. These involve preparation of a scoring stencil, setting up and balancing the circuits in the machine, scanning and cleaning the answer sheets, and scoring and checking. The first step in scoring any tests is the preparation of a scoring stencil. The stencils are made of a lightweight, hard cardboard. Pre-punched stencils are provided by the publishers of many recent tests, where the test has been designed for machine scoring. For other tests, and of course for tests that have been locally prepared, stencil blanks must be obtained from IBM and punched locally. For a test in which the score is the number right minus some fraction of the number wrong, the scoring stencil needs to have only the right answers punched out. When some responses

are counted neither right nor wrong (i.e., receive zero credit), an additional "item elimination" stencil is required. Once a stencil has been prepared it can be used for a long time.

The prepared stencil is placed in the machine, and a check answer sheet is inserted. The check answer sheet contains a known number of right and wrong answers. Using the check answer sheet, the rheostats are adjusted until the machine gives the correct reading for rights alone, wrongs alone, and the combination of rights and wrongs. Several check sheets varying widely in number of rights and of wrongs are used to make sure that the machine is giving accurate readings throughout its complete range.

Before scoring, it is usually necessary to scan the answer sheets for weak marks, incomplete erasures, stray markings, and double markings. Careful instruction and demonstration of the use of the answer sheet prior to testing minimize difficulties on this score. However, occasional papers are found that are not well marked. Unfortunately, the Test Scoring Machine has no way of discriminating between intentional and unintentional marks. A weakly marked answer may or may not register; a poorly erased response or a stray mark may or may not register. For accurate scoring, these points need to be checked.

When the answer sheets have been cleaned up, they are put in the machine one at a time and scored. A switch is pressed to bring the contact brushes firmly against the answer sheet for scoring, and it is pressed again to release the answer sheet, which then drops into a holder in the base of the machine.

From time to time during the scoring it is desirable to recheck the setting of the machine rheostats by means of the check sheets. When a set of papers has been scored, it is a very desirable precaution to rescore them as a check upon the whole operation. Preferably, this should be done by a different operator at a different time. It should be based on an independent setting up of the machine. Ideally, the check scoring may be done on a different copy of the machine.

The chief factor against using the machine is cost. There is first the monthly rental on the equipment, a basic cost of \$40 for a machine. Second, answer sheets for the machine are

protected by patent and must be bought from IBM. The cost varies considerably, depending on the quantity ordered. In 1948 the following prices per 1000 answer sheets were quoted:

	<i>Printed One Side</i>	<i>Printed Two Sides</i>
Order of 500	\$13.00	\$19.00
Order of 2,000	9.75	15.50
Order of 5,000	8.50	14.00
Order of 25,000	6.25	9.50

The scoring machine is most efficient when the job involves the scoring of a large number of papers all of the same kind. If a number of different tests are being used, but only a small number of each are to be scored at a given time, the machine would hardly prove economical. The time required to put in a new scoring stencil, reset the rheostats, and check the settings with standard check answer sheets would use up much of the time gained from faster scoring once the scoring was under way. In a testing program in which the day-to-day populations are small, therefore, even though a number of tests have been given to each person, machine scoring has little advantage. Only if it were possible to accumulate several days' answer sheets would a saving of time be effected. In figuring the economy of machine scoring, some time allowance must be made for examining and erasing stray marks on the answer sheet, because the scoring machine is quite responsive to even rather small scattered markings.

The use of the Test Scoring Machine is, of course, further limited by the pattern of the answer sheet. The machine has 750 contact brushes, arranged in 10 blocks, 5 by 15. This makes it possible to respond to 50 fifteen-choice items or 150 five-choice items on a single side of a page. Other numbers of response alternatives and of test items are also possible. However, the response is limited to that of selecting from among a given set of response alternatives. The answer form can be adapted only with difficulty to the type of item in which the subject himself produces the answer. The rigid spatial pattern of places for marking answers is also sometimes an inconvenient restriction on the form of a test.



### *Procedures for getting weighted composite scores*

If a battery of tests is being used to give a composite score, predicting aptitude for a particular occupation, efficient and accurate routines must be worked out for combining the scores on the single tests into the composite score. This operation involves a good deal of arithmetic and can become both time-consuming and a source of error. Three alternate procedures will be discussed briefly. The first uses only a calculating machine, the second involves the IBM Test Scoring Machine, and the third requires a complete IBM punched card installation.

In any small personnel selection project, and particularly where scoring is being done manually, manual routines for computing composite scores are the most appropriate. Each score is multiplied by its appropriate weighting factor for the job whose composite score is currently being computed, and the products are summed to give the composite score. This operation can be facilitated if a convenient blank is prepared for entering the test scores. One column of the blank should list the names of the tests, and a following column should provide place for entering the test scores for the examinee. Space should be provided for name and any other needed identifying data. Either on the blank or on separate weighting stencils, the weights to be used in computing the composite score for job A, job B, job C, etc., should be listed, their position corresponding to that of the tests.

Computation of the sum of products can be carried out on any available type of computing machine. An experienced Comp-tometer operator can carry out this type of summing very rapidly. A cumulative sum can also be obtained on a Monroe, Marchant, or other make of computing machine. Whatever machine is used, these composites should invariably be checked by an independent computer. There is no other check for errors, and occasional errors will creep in with even the most skilled computer. Discrepancies between original and check computer should, of course, be especially carefully checked.

If the IBM Test Scoring Machine is being used to score the separate tests, it is possible to use this machine also for computing weighted scores. This can be done by using a device called

an *aggregate weighting board*. The aggregate weighting board is used with special record sheets called aggregate weighting sheets. These sheets have 30 positions, arranged in 10 rows of 3, in which it is possible to represent score on a particular test. Score is represented by marking with an electrographic pencil a pair of lines whose length corresponds to the score on the test. One line represents digits and the other tens. Thus, a score of 35 is represented by a line extending through three steps in the tens position and one extending through five steps in the digits position. Each test is assigned a particular one of the 30 positions on the record sheet and is represented in that same place for each individual tested.

Weighting is accomplished by a plug board which makes it possible to assign a weight having any integral value from 0 to 20 to the tests in a particular row. Different weights may be assigned for each row, but all the tests in the same row must receive the same weight. When the plugs have been inserted in the aggregate weighting board in the appropriate positions and the board inserted in place of a scoring stencil in the Test Scoring Machine, the record sheet for a particular individual may be inserted, and a weighted sum read off directly from the dial. Because of the limitations of the dial of the machine, this sum must be limited to the range 0-100. An additional over-all weighting constant is therefore required in the setting of the rheostats of the machine to bring the composite scores within the desired range. The effective range of score values becomes perhaps 50 or 60 points. Since on most tests even those individuals who do poorly receive appreciable scores, their weighted composite will have a positive value of moderate size, and the range of composite score values from 0 to 40 or 50 is not likely to appear.

The aggregate weighting procedure is very rapid, once the test scores have been marked on the aggregate weighting sheet. Entering these and checking the entries is a somewhat more laborious process than entering scores on a record sheet for use with a computing machine. There are certain mechanical limitations in this procedure which merit our consideration. The first limitation is that the machine provides no direct technique

for giving negative weight to a test. The only way this can be done is to subtract the score from a constant and enter these inverse values on the aggregate weighting blank. If a test is to be given a positive weight for one job specialty and a negative one for another, it must be treated essentially as two distinct tests and entered on the aggregate weighting sheet once as the actual score and once as the inverse score. This is time-consuming and also uses up the space on the aggregate weighting blank.

The second limitation of the aggregate weighting board is that it provides for only 10 different weights for tests. Different weights can be assigned to each of the 10 different rows on the blank, but if more than 10 test variables are used it is necessary to enter two or three variables in the same row. These must then receive the same weight *for every job category*. In a program making use of an extensive battery of test variables, and particularly in a battery from which a number of composite scores are being derived in order to estimate aptitude for a number of different job categories, this limitation may be a very real one. The flexibility of the weighting system may be considerably reduced by the restricted number of weighting positions in the aggregate weighting board.<sup>2</sup>

The economy of the aggregate weighting board, as opposed to the calculating machine, as a device for computing composite scores is once again a function both of the total flow of records to be processed and of the number for which a particular calculation can be done at one time. If after the scoring machine and weighting board have been set up and checked it is possible to run through two or three hundred papers before the machine must be reset, the economy of time will probably be substantial. But no economy could be achieved in getting ten different composite scores for each of twenty individuals, because this operation would necessitate a resetting of the machine for each new composite score.

<sup>2</sup> Some increase in the variety of weighting patterns may be obtained by masking out certain of the 30 test positions on the board with a non-conducting masking tape. It is thus possible to give one test on a certain row of the blank zero weight for a particular job category while giving a positive weight to the other test or tests.

Where the flow of testing is very large and where punched card records are being used as the basis of the permanent record system and for research analyses, so that an installation of IBM equipment is already available, consideration may well be given to using punched card methods as a basis for computing composite scores. The particular procedures depend on the particular units of equipment which are available. One procedure is based on the machine known as the *multiplier*. This machine will multiply the scores for each individual by the specified test weights and enter the products on the IBM card. The products can then be summed in a separate operation. The capacity of the multiplier is limited to two or three variables at a time, so that with a large battery of tests a number of runs is required to get all the products and to sum them.

Another procedure for getting composite scores makes use of the *reproducer*. With this procedure, the score for each different test must be entered on a separate IBM card. A set of master cards is made for the first test, one card for each possible score. On the master cards are punched not only the raw test score, but the score multiplied by each one of the weights which is being used for that test for the different job categories. The cards for the group of subjects are sorted so that they line up behind the master card which has the same score that they do, and the reproducer "gang punches" the appropriate products of weight times score into each single card. The operation is repeated for each one of the tests. It is then necessary to take all the cards for each individual, one for each test, and sum the products on these in order to arrive at the composite scores. This can be done on a *tabulator* and the final results can be punched on summary cards.

The above description is very abbreviated and will have little meaning to a person who is not familiar with IBM equipment. However, it may suggest to a person working with such equipment ways in which the equipment can also be used to handle these computations. The procedures are sufficiently complex so that they are economical only when the flow of persons tested is decidedly large and the records of the separate test scores have to be punched on cards in any case for permanent record and future research use.



### *Procedures for auditing and checking*

In any testing program that prides itself on accuracy in the scores and final aptitude ratings which it reports, auditing and checking procedures must become a standard part of the operating procedures. Two types of checks should be included. These are (1) routine independent repetition of any operations based primarily on the human machine and (2) complete spot checks of the whole process of arriving at aptitude estimates for sample individuals. The first checks accuracy in those operations in which human fallibility is most pronounced, and the second reviews, by sampling, the accuracy of the whole scoring and computation process and locates systematic errors which may have served to invalidate a whole group of scores.

Just which operations should be checked in full depends on the available facilities and the extent to which accuracy in handling each case is a critical consideration. An industry might feel, for example, that it could afford to work from erroneous scores in a small percentage of cases, whereas an agency examining for civil service appointments might consider scrupulous accuracy of score for each individual a critical matter. The operations that should most certainly be checked in full are those into which the human factor enters each time the operation is performed. Outstanding among these is any form of computation, especially mental computation. Second comes any form of copying or entering of a score. Probably somewhat less critical, because errors are likely to be smaller in size, is any hand scoring and counting of tallies. Machine scoring or machine weighting of scores in a score composite is probably still less urgently in need of a complete check. However, even in machine scoring occasional errors are found, and a meticulous testing program may well demand complete checking of even these operations.

Spot checking is a different type of operation and is designed to serve a rather different purpose from complete repetition of particular operations. In spot checking, a small sample of individuals, perhaps two or three out of every hundred tested, is selected for detailed checking of the whole sequence of opera-



tions applied to the test results. Each procedure is carried out in so far as possible by a different method from that originally used, so that any constant error which may have crept into the procedures applied to a particular batch of tests may be detected. Thus, in a program using the Test Scoring Machine for routine scoring and computation of composite scores, the spot check would use hand scoring for obtaining test scores and calculation with a computing machine to determine the composite scores. This auditing should preferably be carried out by personnel different from those involved in routine scoring operations. Test keys for auditing should be kept in a separate place, and an entirely independent set of reference materials should be used in determining the weights to be applied to test scores when they are combined, and so forth.

All the procedures for auditing routine scoring and computing procedures are designed with the goal of having them as distinct from the routine procedures as possible. This is to minimize the possibility that any error which crept into the routine procedures may also affect the spot check. Some types of error which spot checking may catch are the following:

1. Errors in adjusting the Test Scoring Machine, leading to a constant error in score throughout a group of papers.
2. Use of the wrong scoring key for a test.
3. Use of the wrong scoring formula for a test.
4. Use of the wrong set of test weights for a job category.
5. Use of the wrong multiplying constant when the Test Scoring Machine is used to yield composite scores.
6. Use of the wrong check sheets in adjusting the Test Scoring Machine, either for scoring or for computing weighted composite scores.

These types of error may seem impossible to the person who has had no experience with a large-scale testing program. The experienced worker knows that they do in fact occur. Even though this occurrence is quite infrequent, it can still be serious, because the resulting errors may be large ones and they may affect a whole set of scores. Spot checking with different scoring and weighting procedures does not guarantee to catch all errors, but it provides the basis for locating many of them.

## ORGANIZATION OF REPORTS AND RECORDS

When testing has been completed for a group of individuals, the tests have been scored, and the scores combined into weighted composites, the final stage of preparing the results for use consists of arranging them in proper form and making them available to the ultimate user. In this connection, we must distinguish between two types of users. There is the immediate user, who must have the test record as a basis for making an administrative decision as to whether a given applicant is to be employed or in what capacity he is to be used. There is also the ultimate user, who is going to use the results at some later date, usually for some research project, such as validation of the test scores against later measures of success on the job. Sometimes the record is needed for some later administrative decision about promotion or special assignment. The important factors in these two uses of test records are somewhat different and merit separate consideration.

For immediate administrative use, the important considerations are (1) prompt availability of the report, (2) maximum simplicity of the record consistent with providing the necessary information, and (3) ease of locating the record for particular individuals. For immediate use, some type of daily roster of test results may prove satisfactory. If administrative decisions are handled on a day-to-day basis, keeping the records from each day's testing together may prove an efficient basis for organization of the immediate report. This roster should include only the information which is to be considered in arriving at the administrative decision. Thus, while composite aptitude scores for each job specialty should certainly be included, scores on separate tests of a battery may well be omitted. Such a roster can be prepared from the composite aptitude score records of each day's testing, as soon as they have been checked, and made available to the person responsible for making an administrative decision on each case.

The essential qualities in the permanent system of records which are to be referred to at a later date primarily for research studies are (1) completeness, (2) ease and accuracy with which

the records can be collated with later records of criterion performance, and (3) ease with which statistical analyses of the records can be carried out. In contrast with a report to be used for an administrative action, research records should be as complete as time and resources permit. All available information should be retained, since it is almost impossible to tell in advance just which items may be needed. Furthermore, the records should be so arranged and organized that the correspondence between particular individuals in the record file and the individuals for whom later data are obtained can readily be established. Finally, records should be in convenient form either for manual or machine computation of needed statistics. These last two points of ease of collating and ease of computation merit somewhat fuller exploration.

The first point with regard to permanent test records is that they should be in such a form that it is easy to collate them with other information. The collating operation most frequently needed is to bring together the test records and criterion information for the same group of individuals. Thus, some time after the tests have been given, records are obtained of grades in or graduation from a course of training, or ratings or performance records are obtained for the workers in a particular factory. Working back from this criterion information, it is then necessary to bring the test and criterion records together so that computations can be made involving both of them. The collating operation involves three steps: (1) locating the earlier record, (2) positive identification of the individual, and (3) bringing the data physically together for use in computing. A good organization of the records of testing will facilitate each of these steps.

For the greatest ease in finding a test record, it should be necessary to search only a single file which is arranged in accordance with a single system. To search through two files each containing 10,000 entries will take substantially longer than to search one file with 20,000 entries. Each further increase in the number of places in which one must look represents a further increase in the work involved. The system of arrangement for the record file should be one which is readily applied also to the criterion records.

For most types of hand collating, the most practical arrangement for finding records, either for collating with a criterion record or for other uses, is an alphabetical file by name. The use of name has advantages over any numerical coding system because of the meaningful organization of the letters as they form the name and because of long familiarity with names. The file should almost certainly be a card file, with each individual represented by a separate card. This makes it possible to pull any desired records out of the file and also to add new records to it continuously.

Where the program involves only a single category of placement, a single file should suffice. However, it may be desirable to maintain an active and an inactive file, removing to the inactive file those cases who, because of the length of time since testing, are very unlikely to appear in current research studies. Similarly, it may be desirable to keep in a separate inactive file (or possibly not to file at all) the results for rejected applicants who have not been given employment. The goal is to keep the active file complete, so that it will include all but a very few of the individuals who might be found in a current research study, and yet to keep it as compact as possible.

Where two or more job categories are under study in the same program, the problem of efficient organization of files becomes more complex. If it is impossible to tell at the time of testing or shortly thereafter into which job category an individual will fall, or, if transfers from job to job happen fairly frequently, it is desirable to keep the records for all jobs in a single file. Separate files for the separate jobs are not practical in this case. However, if assignments are known at the time of testing and there is little or no interchange of personnel between different jobs, a separate file for each job is advantageous. The separate files will be smaller and will reduce the population to be searched in matching the test records with the criterion data.

In many large-scale personnel research projects, it is advantageous to use IBM punched card equipment for record keeping and for research analyses. Collating is then usually carried out on the machine specially constructed for that purpose and called a *collator*. This machine matches cards from a pack containing criterion information against the total file of test record cards



and picks out from the larger testing file those cases which occur in the criterion file.

If the collator is to be used, it is no longer advantageous to file and match by name. This is due to the fact that the names of two different individuals may be either completely the same or the same for a great many letters. One is limited in the number of letters of a name on which it is practical to match when collating. Also machine alphabetizing and merging of two files becomes very time-consuming. Furthermore, matching should be uniquely correct. There should be no possibility of coding two individuals in the same way. For use with punched card records, the best basis for collating seems to be a unique number designating each individual.

In the armed forces, a unique coding was provided by a serial number assigned to each man at the time he entered the service. In civilian life, an equivalent which could often be used is social security number. In particular organizations, there may be some other unique coding number which stays with an individual and identifies him throughout his career in that organization. In each of these examples, a particular number sequence always designates a unique individual. Barring copying and similar errors, no mistake is possible in matching two sets of data for the same individual. For machine collating, therefore, files can often advantageously be organized by some type of serial number which stays permanently with the individual.

The possibility of finding two individuals with identical names, where name is being used as the basis for matching, or of a copying or reporting error where a serial number is being used, points up the necessity of a systematic check upon the correctness of matching. After a pair of records has been matched on one basis, it is always desirable to have another independent basis for checking the correctness of the matching. For this purpose, it is always desirable to include at least two or three clues to identification in any record card. If the file is an alphabetical file by name, a check clue upon correctness of matching may be provided by date of birth, date of employment, or by social security number. Where a numerical coding is the basis of filing and matching, a good check is provided by name. This secondary cue should be referred to routinely in cases of hand



collating. It also provides a basis for tracing down errors and reconciling discrepancies in machine collating.

Collating sounds like a very simple operation, and basically it is. But in a testing program of any size, it can become very time-consuming, especially where the effort is made to carry it out with a high degree of completeness and accuracy. If the test records are not systematically organized, the task can become a hopeless one. Keeping test results recorded and filed accurately and currently represents a fairly substantial undertaking for which provision must be made in the budget of any research program.

Once the sets of data have been collated and the accuracy of the matching has been checked, the test and criterion scores must be used together for computation. How the two sets of data shall be organized for efficient computing depends again on whether the original records are in the form of a written record or of a punched card record. If the records are written, some provision should probably be made on the test record card for entering later criterion data. If certain spaces are set aside for criterion data, these data can be entered directly on the card which contains the test results. This card can then serve as the master source, either for punching the scores for the group of cases being studied into punched cards for analysis or for hand tallying and calculation. If test and criterion records are already on IBM punched cards, then the data from one set can be reproduced into the other, or selected data from each can be reproduced in a new work deck of cards by using the *reproducer*. The cards with the combined data then serve for computations.

We may now consider the arrangement of a test record card. Some of the problems of a written and a punched record are different and require separate consideration. Problems which we shall consider are (1) the location of identifying information, (2) the location of test results, (3) provision of space for criterion and other research data, (4) dealing with changes in the test battery, and (5) the number of digits to be used in reporting scores.

On any card used for hand collating, identifying information should be placed at the top of the card. The information to be

used as the primary basis of matching should be at the left, since the top left is the place which we normally observe first. The information to be used as a check upon the matching can well be at the top to the right. It often pays to have this identifying information reproduced on the face of the card even when punched cards are being used and most of the collating and other statistical operations are being done by machine.

With a written record system, test results spread across the body of the card. In any but very small-scale projects, it usually is desirable to use a printed or photo-offset form, so that each score to be recorded can be designated and will appear in its own particular space. Ruling may be used to separate the different scores. The arrangement is a compromise between the desire not to crowd the scores together on the one hand, and the desire to provide space on the card for as much supplementary research information as possible. The usefulness of the card is increased if it provides space for a number of items of research data—scores on research tests, criterion records, and the like.

Any record system must face the problem of change. As time goes by and additional data are accumulated, it will seem desirable from time to time to make changes in the battery of tests which are being used for selection or classification. Frequent changes make the use of the records for research analyses quite difficult. This is true particularly if machine methods of collating and tabulating are being used. The efficiency of these methods depends on uniformity in the records, so that the same items of information are entered in the same place on the card for every individual. The whole tabulating job can then be handled in a routine and mechanical manner. Every time test procedures are changed, with some tests being dropped and others being changed or added, it means that the records must be treated in fractions, each fraction being handled slightly differently. This is true of hand tabulation also, but manual tabulating routines are more flexible and adjust more readily to shifts in the data.

The record system should be set up so that changes in the test battery will disturb the tabulation of results as little as possible. First of all, every record card should be clearly coded to desig-

nate which battery was in use when that individual was tested. In the second place, the location of the score of a particular test on the card should be changed as infrequently as possible. That is, the tests which remain unchanged from one battery to the next should also remain in the same position on the card. If the space on the card permits, some blank spaces should be reserved and these should be used for new tests when the battery is changed. The positions of scores of tests which are dropped from the battery should be left blank. If it is possible to do this through two or three changes in test battery, then in both hand and machine tabulation it will be more feasible to work with data from those several batteries simultaneously. There will be no ambiguity about the score occurring in a particular position. It will refer to a particular test, and if that test was not given the position will be blank.

The problems of space, especially on a punched card, and of efficiency in computation raise the question of the number of digits to be retained in test scores. In particular, these problems raise the question of the desirability of converting all raw scores into single-digit scores on a standard score scale. Single-digit scores make for great simplicity and economy of tabulation, especially with machine tabulation, where the number of tabulator runs necessary to handle complex correlational data and similar problems is greatly reduced. Again, particularly in machine record methods, the use of single-digit scores increases substantially the number of items of information that can be included in one punched card. Each test then requires only one column on the card, and many tests can be included without exceeding the 80-column capacity of the card. The use of single-digit scores is to be recommended especially for programs in which the testing battery includes a large number of separate tests. In this case, the increased capacity of record card and increased efficiency of tabulation become very important, whereas the small loss of accuracy due to the grouping of scores into nine or ten categories on each test becomes a very minor matter. The coarseness of grouping on any single test is compensated for by the number of different tests. When the tests are few in number, any advantages in single-digit scores are minor, and these advantages probably do not compensate for

the loss of information through grouping and for the work involved in converting the raw scores to single-digit scores.

### CONCLUDING STATEMENT

In this chapter, an effort has been made to suggest some of the administrative problems in a large-scale personnel testing program. In part, these problems refer primarily to day-to-day operations; in part they occur in research with the test results. Problems have been identified and some suggestions made for procedures in giving, scoring, checking, and weighting the results from tests, and in organizing test records for routine and for research use. The discussions have necessarily been in rather general terms, since details of type of testing battery, size of flow of subjects, conditions of scheduling, deadlines for providing test results, and other unique administrative features will determine the pattern of procedures in any specific testing program. In any event, it will pay to devote a good deal of thought and care to the planning of these administrative routines, if an efficient testing program is to result.

## *Administrative Problems in Using the Results of an Aptitude Testing Program*

After a battery of aptitude tests has been administered and scored, and after the results from the single tests have been combined with appropriate weights to yield the best prediction of success in one or more job specialties, then the test results must be put to use to select or reject specific job applicants. Administrative routines must be established for translating the composite aptitude score for individual  $X$  into an administrative decision that individual  $X$  shall be hired, or for translating a set of composite aptitude scores in jobs  $a, b, c, \dots, k$  for individual  $Y$  into a decision that individual  $Y$  shall be assigned to job  $i$  rather than some other job.

### TYPES OF NECESSARY ADMINISTRATIVE DECISION

The pattern of administrative decision required for a given case may be any one of three types:

1. Selection-rejection.
2. Multiple selection-rejection, or multiple qualification-disqualification.
3. Classification.

In selection-rejection, the individual is being considered for a single job category, and the decision is simply that he shall (or shall not) be accepted for that job category. The practical problem is to fill vacancies in that particular job specialty as they occur with the most suitable individuals from among the group of applicants. This should be done promptly and yet without the bother and expense of maintaining an extensive pool of accepted applicants for whom no work is available.



In multiple selection-rejection, the individual is considered a candidate for a number of job categories, but the action to be taken with regard to him is to be merely negative. That is, a decision is to be made as to the job categories for which he *is not qualified*. No positive action is taken from among the ones for which he *is* qualified. The administrative problem is to have sufficient numbers qualified in each job category so that the needs in that category can be met at the same time that the needs in other categories are met. Thus, if the weekly demand is for 50 auto mechanics and 30 apprentice machinists, we must be sure that our procedures give us both (1) a total weekly yield of at least 80, and (2) specific yields of at least 50 qualified as mechanic and at least 30 qualified as machinist. Due to the overlapping of qualifications, it is quite possible that 30 of the qualified mechanics would also qualify as machinists. However, 50 men could not fill 80 jobs. We must consider not only single job categories but combinations of categories. Again, the flow must be sufficient to meet needs and commitments, but not so great as to result in pools of personnel for whom no jobs are available.

Classification involves the assignment of each individual who is accepted to a specific one of a number of job specialties. Assignments must match the demands and opportunities, and at the same time the available resources of job applicants must be used to the best possible advantage. The administrator faces a complex matching problem involving men and jobs, in which an effort is made to use aptitude information in a positive manner, so that each individual's areas of specially high aptitude are taken into consideration in affecting his assignment. The process is complicated by the fact that assignments of individuals interact. If individuals *A*, *B*, and *C* are assigned to job *k*, and there are only three vacancies in that job category, then individual *D* cannot be assigned to that category of job. In assigning any single individual, attention must at the same time be given to the pool of aptitudes represented by the other members of the group.

This discussion has indicated that aptitude test scores may be used either as a hurdle or as a quantitative indicator of success. In the former case the score is used in an all-or-none manner to separate the sheep from the goats, those who shall be accepted from those who shall be rejected. In the second case, the score

is treated as a continuous variable and as an indicator of the probable degree of success in the job in question. For the routine administrative use of test results, use as a qualifying hurdle is much the simpler undertaking. Once the minimum qualifying score for a job category has been determined, the limits of administrative action for a given individual are immediately specified. He either is qualified or he is not, for each job category. He may be assigned only to a job category for which he is qualified. The possibilities are rigidly and objectively specified, and job assignment can then be made by any competent clerk.

When tests are treated as continuous quantitative variables and indicators of probable success, the administrative problem becomes a good deal more complex. It becomes necessary to assess the potentialities of a candidate for each job category and to arrive at a decision as to which assignment represents the most advantageous use of his potentialities. The use to be made of one man depends on the other men for whom an assignment must also be found. If assignments are to be made to best advantage, either a rather complex set of ground rules is necessary to reconcile aptitudes and job vacancies, or a high standard of professional judgment is required on the part of the person making the assignments. The possibility of using scores in a positive and constructive fashion for the purpose of classifying job applicants can be gained only at some sacrifice of administrative convenience.

### FACTORS INFLUENCING THE ADMINISTRATIVE PATTERN

Reaching an administrative decision with regard to assignment is a problem of reconciling supply, demand, time limits, and aptitude ratings. We shall need to consider how these factors interact.

In any personnel selection enterprise there is a certain available supply of job applicants, from which the needs of the employing organization must be met. This supply is not fixed, since it depends on the extent and effectiveness of the personnel recruiting program. With a more intensive recruiting campaign

or an increase in the inducements offered to applicants, the supply can be stepped up. As competing opportunities are made more attractive, the supply will decline. Unless the supply of applicants for a particular job is in excess of the need for persons in the job, a selection program is largely futile. Unless one can afford to reject some of the applicants for a particular job, there is little use in testing them. Of course, one may wish to maintain minimum standards of aptitude or proficiency even though it is impossible to fill all the existing job vacancies. Test results may be used for this purpose. But test results and other selection procedures function most effectively when the supply of job applicants is substantially in excess of the number of placements to be made in the job.

The limitations which supply puts on the use of selection procedures emphasize the importance of the personnel recruiting program. The characteristics of the supply of job applicants, in terms of both quantity and quality, set the limits of what a selection program can achieve. Even a perfect set of selection procedures can only pick the best from what is available. If recruiting has been ineffective, the best may be none too good.

Where the project involves classification rather than simple selection and each applicant may be considered a candidate for more than one job category, the supply-demand problem becomes rather more complex. Numerically, it is the over-all supply which is critical. Unless recruiting is hopelessly inadequate, the number of applicants will be in excess of that needed to fill vacancies in any one job category. Satisfactory recruiting must supply more than enough for the needs of all the job categories. In addition, recruiting must supply a population of job applicants whose qualities match the needs of the several jobs. Since there are always problems of best utilization of personnel in a classification situation, even though there is no over-all excess of job applicants, test results serve a vital function in the classification of personnel even when applicants are in short supply. Measures of aptitude have a valuable function in helping to achieve the most advantageous allocation of the existing supply of applicants to the existing job vacancies.

Implied in the term *supply* as we have been using it is the acceptance of the job by the applicant. Thus, if only 50 per cent

of those who originally inquire about a job will finally accept it if it is offered to them, then a weekly flow of 200 persons tested represents an effective weekly supply of 100. Furthermore, the qualifying score will have to be set at a point to qualify approximately twice as many individuals as will ultimately be needed for employment. Variation from day to day in rate of accepting employment will introduce one more source of variability. In any situation except a coercive wartime military situation, this problem of the final acceptance rates of offered employment will have to be reckoned with.

Time is another vital factor in determining administrative routines in personnel assignment. On the one hand, there is a limit on the time that can elapse before a decision must be made with regard to a particular man; on the other, there is a time limit within which a particular job or quota of job assignments must be filled. One cannot keep a job applicant waiting indefinitely. At some point he must either be offered a job or be told that none is available for him. In some cases, it may be possible to keep him waiting for several days or weeks, but in others it may be necessary to come to a decision while he is physically present and making his application. It depends on the type of position for which application is being made, the press of other job opportunities for the applicant, and the nature of the contact which the organization has with the applicant. At the same time, from the point of view of the employing organization, there may be certain deadlines which must be met. Perhaps the most definite of these is a specified date for the beginning of a program of instruction. However, there will always be some pressure from the employing agency for promptness in making job assignments.

The longer the time period which may elapse before an administrative decision must be made about a particular individual, the more completely it is possible to utilize aptitude information in the selection and classification of personnel. It is then possible to accumulate data on a large group of individuals, and to make selections and assignments from this large group. Day-to-day fluctuations both in the applicants and jobs to be filled are evened out. Each decision with regard to assignment can be made with reference to a large group of candidates, and



the probability of finding within that large group individuals with the specific qualities needed in a particular job is relatively good. When a decision concerning a particular individual must be made immediately, the momentary supply-demand situation may influence that decision as much as or more than aptitude.

The press for immediate decision was one of the most severely limiting factors in army personnel utilization during the last war. Men came to reception centers, where they were subjected to certain aptitude tests and a classification interview. On the basis of test scores and data on educational and vocational history, assignments were made to replacement training centers or to special schools. However, this testing, interviewing, and assignment had to be completed within two or three days. Furthermore, assignments had to be made continuously to meet current quota demands. If a quota of 100 entrants into radio mechanic's school had to be met by Friday, then Wednesday and Thursday had to yield a good crop of future radio mechanics, no matter how unpromising the raw material, and a top-ranking radio mechanic who showed up on Saturday had no prospect of getting in. There was limited provision for pooling promising recruits against anticipated future quotas. Immediate demands became a very compelling factor in personnel classification.

### DAILY QUOTA VERSUS PREDICTED YIELD AS ADMINISTRATIVE PATTERNS

There are two distinct approaches to adapting the aptitude score information to the limitations of supply, demand, and time for reaching a decision. In order to identify these, we may label them the *daily quota* method and the *predicted yield* method. We shall elaborate these procedures first as they apply to a problem of simple selection and then as they apply to a problem of classification.

In applying the *daily quota* method to a problem of simple selection for a single job category, the procedure is to arrange all those tested on a given day (or within some other unit of time that is administratively convenient) in order in terms of the score which represents the final composite estimate of their aptitude. The number required to meet the current quota of positions to be



filled is counted off, starting at the top of the list. Thus, if 100 individuals have been tested in a given unit of time and 25 positions are to be filled at that time, the positions are offered to the top 25 in order of aptitude. If 80 are tested during the next time period and a quota of 60 comes in to be filled at that time, the top 60 of the 80 are chosen. The score which separates those who are accepted from those who are rejected fluctuates with supply and demand, going up when applicants are many and jobs are few and down when applicants are few and jobs are many.

This procedure is administratively very simple, since it is necessary only to take account of the immediate group of applicants and the immediate quota of jobs. The pattern is cut to the cloth currently at hand. However, if the system operates in terms of such small time units that there is appreciable fluctuation in supply and demand it becomes both unfair from the viewpoint of the applicant and inefficient from the standpoint of the employer. On a day when the momentary demand is light, individuals will be rejected who have substantially higher aptitude for the job than those accepted a few days later when a new quota has to be met. It is this unevenness of standards which the method of *predicted yield* undertakes to overcome.

In the method of *predicted yield* an effort is made to forecast the general trend both of supply and demand. The essential facts are (1) the expected number of daily (weekly, monthly, or other convenient unit of time) applicants for the job, (2) the expected distribution of scores on the composite aptitude measure, and (3) the expected daily (weekly, monthly) quota of positions to be filled. The estimates for the future have their basis, of course, in experience in the past. The first essential for any forecast of future supply, future demand, and future distribution of aptitude scores in the population of applicants is accurate information as to what the supply, demand, and aptitude distribution have been to date. These statistics from the past can then be adjusted to take account of any known changed conditions in the future. Thus, it might be planned to expand operations of a particular type, and this expansion might necessitate a 50 per cent increase in the placements in a particular job category. Or the progressive exhaustion of the supply of qualified appli-

cants might indicate a gradual lowering of the average aptitude in those who would be available for testing. On the basis of the best estimate of the total number of applicants, the distribution of aptitude in the applicant group, and the number needed for employment, a qualifying score can be set which will qualify the desired number of individuals. Thus, if there are 150 applicants a month for a type of position and 100 vacancies to be filled each month, the minimum score chosen for that job is the one which estimates show can be reached by two-thirds of the group of applicants. The primary basis for choosing this score is the distribution of scores for past groups of applicants who have taken the test battery.

The chief advantage of the *predicted yield* method is that it sets a stable and uniform minimum qualifying score. It holds a fixed standard, despite short time fluctuations in supply and demand, and so guarantees the highest average level of aptitude in those accepted for the positions.

The disadvantages of the method are twofold. In the first place, it does not give an exact day-to-day correspondence between positions to be filled and persons declared qualified to fill them. At one point there may be a few more vacancies than there are applicants who have been tested and who meet the aptitude standard for employment. A few days or weeks later there may be an excess of qualified applicants over the number of positions to be filled. This necessitates some procedure for taking up the slack, some way of pooling a reserve of qualified applicants.

The second problem of this approach is one inherent in any problem of forecasting. Our forecast may turn out not to be accurate. We may miscalculate with regard to either our supply, our demand, or our aptitude score distribution. If this happens, the yield of qualified applicants will not match the need for them and we will face a systematic excess or deficit. The only way to forestall this occurrence is to maintain continuous records of the supply-demand balance. As soon as it is clear that the situation is getting systematically out of balance, an adjustment will have to be made in the minimum qualifying score so as to restore the balance. If the supply is in excess the qualifying score will have to be raised, whereas if the demand is in excess

the qualifying score will have to be lowered. (Of course, a preferable adjustment may be to intensify or reduce the recruiting, so that the supply is modified rather than the qualifying score.) These problems of providing for temporary surpluses and deficits and of adjusting for inaccuracies of one's forecast make the administration of this procedure somewhat complex.

A combination procedure is possible and may sometimes be useful. This procedure is to set a fixed minimum qualifying score, but to set it low enough so that the number qualified will in general be in excess of the demand. Beyond this minimum score, positive selection can operate each day, taking first those with the highest aptitude. The minimum qualifying score protects the organization from accepting inferior personnel at moments of peak demand or short supply, and the general excess of supply over demand makes it unnecessary to pool any substantial number of individuals against anticipated future needs.

The daily quota and predicted yield approaches operate also in the classification situation, though the lack of a clear, mathematically best solution to the classification problem makes it difficult to state in unambiguous terms how any classification system will operate in practice. The *daily quota* approach to classification means that we take on the one hand the group of individuals tested during a particular day (or other unit of time), with their several aptitude scores. On the other hand we line up the number of appointments to be made into each of the job categories for which the applicants are being considered. Taking account of the relative importance of the several jobs, we sort and juggle until the required number is obtained for each job and the aptitude scores in the different categories appear to be about as high as we can get them. The operation is strictly a trial-and-error one of doing the best one can with the current set of quotas to be filled and the current supply of applicants with their several aptitudes.

Perhaps even more than in the case of selection, classification by this procedure is at the mercy of the day-to-day fluctuations of quotas. If one day provides a large quota of clerks, whereas the next day provides a large quota of mechanics, we may find good potential mechanics being made into clerks one day and good potential clerks into mechanics the next.

The application of the method of *predicted yield* to a classification problem also becomes somewhat more involved than in the case of selection. The task is to devise a set of working rules for classification which will tend to get the best possible individuals in each job category and which will also yield an output in each category that corresponds to the vacancies available in that category. Thus, if our predicted average need for pilots, navigators, and bombardiers is respectively 800, 100, and 100 per month, a satisfactory set of rules for classification must on the average yield those numbers. Again, the essential data are the predicted supply of applicants, the predicted demand for personnel in each of the categories, and the predicted aptitude score distributions. However, it is now the joint distribution of all of the aptitudes which is important. In estimating yield, it is important not merely to know how many men will fall above score  $X_i$  for pilot, for example, but also to know how many of those men will at the same time fall below score  $Y_j$  for navigator and score  $Z_k$  for bombardier.

Setting up rules for classification is necessarily a trial-and-error, cut-and-fit proposition. There is no solid analytical, mathematical basis for it. The personnel psychologist sets up an hypothetical set of rules which seem to him likely to give approximately the desired quotas. Thus, the rules for the pilot-navigator-bombardier problem mentioned above might take some such form as the following: "(1) Classify as navigator any man whose navigator aptitude score, expressed in stanine units (a 9-point scale of half standard deviation units) is both (a) higher than his bombardier aptitude score and (b) two points higher than his pilot aptitude score. (2) Classify as bombardier any man whose bombardier aptitude score is both (a) higher than his navigator aptitude score and (b) two points higher than his pilot aptitude score. (3) Classify all other qualified men as pilots." These rules must then be applied to actual or theoretical test data to see how they work out. If composite aptitude scores are available for a large population of cases, those cases may actually be classified according to the set of rules and the actual yield in each category determined empirically. If the yield does not correspond to what the practical situation demands, some adjustment may be made in the rules, and the process repeated. This



may be repeated again and again until the approximation to the quota requirements is sufficiently close.

The above process becomes a very laborious one if more than one or two trials are needed to reach the desired result. Furthermore, the trial-and-adjustment procedure is pretty blind. Operations can be facilitated somewhat if a joint frequency distribution of the several aptitude scores is prepared. This is quite feasible for two or three variables, but for a larger number it would require much work, with a rapid increase in the unwieldiness of the operation for each additional job category. The joint two- or three-dimensional distribution of scores can either be prepared by tallying actual data, or be constructed from theoretical distributions on the assumption of a normal distribution in each variable. If enough cases are available, and particularly if records are available on IBM punched card equipment, the actual joint aptitude score distribution is probably the simplest to prepare. Otherwise, if we know the means, standard deviations, and intercorrelations of the several aptitude scores, it is possible to compute the theoretical proportion of cases in every cell defined by score limits on each one of the several aptitude composites. Given such a joint frequency distribution, it is possible to determine the outcome of any particular set of classification rules by summing up the frequencies in all the cells of the table which are to be classified in a particular way. The effects of a given change in the rules may be anticipated with some accuracy by seeing just which cells are shifted from one category to another. Trying out different sets of rules for classification is greatly facilitated.

Predicted yield as applied to the classification problem again has the advantage that it is uniform, fair, and insensitive to day-to-day fluctuations in populations and in quotas. It also has the advantage that the fixed set of rules operates to reduce the large element of subjective judgment which is otherwise likely to enter into a classification situation. With a complex of jobs and aptitudes to be considered and reconciled, the personnel worker is likely to vary substantially not only from day to day but even from minute to minute in the disposition which he will make of comparable cases. Given a set of explicit rules to work by, this variation should be largely removed. The disadvantages



of the method lie, again, in the administrative complexity, the problems of pooling applicants when classifications and quotas are temporarily out of balance, and the possibility of systematic error in forecasting, leading to a systematic discrepancy between classifications and quota requirements.

The complexities which arise in setting up any fixed system for classifying personnel, at least when the classification is into more than two or three job categories, lead one to question whether the personnel psychologist should take the responsibility for actually carrying out classification and assignment. A multiple selection program based on a variation of the predicted yield technique may prove most satisfactory in practice. The procedure is to set minimum qualifying scores for each job specialty. Each individual's aptitude scores are compared to these minimums to determine for which job or jobs, if any, he is qualified. The several minimum scores are so chosen, with reference to the joint frequency distribution of composite aptitude scores for each of the job specialties, that the total number qualified is sufficient to meet the total demand for personnel and so that the number qualified for each job gives some surplus over the number required for that job. Individuals can then be assigned to any job specialty for which they are found qualified. The assignment of individuals qualified for only one of the jobs will be automatic, and the assignment of those qualified for two or more jobs can be juggled to fit the necessary quotas. This juggling can be done by non-psychologists, since no particular technical training is required for it. This pattern of operation does, in fact, come rather close to the one which was in effect in the AAF air-crew classification program at the end of World War II.

### INCLUSION OF NON-TEST VARIABLES IN THE PREDICTION OF JOB SUCCESS

In most of our discussion of predictors of job success up to this point, the predictor variables have been spoken of as "tests." Of course, all types of data may be used for purposes of prediction, and the predictor variables need not be "tests" in the common meaning of the term. Age, sex, education, and marital

status are items of information which are generally available and sometimes useful. A statement of amount and type of previous job experience is frequently relevant. Expressions of interest or preference for one job rather than another may need to be taken into account. What shall the administrative pattern be for dealing with items of information such as these?

Let us assume that several test scores are being combined into a weighted composite score which is being used as a predictor of job success. Three patterns of use of auxiliary non-test data ordinarily suggest themselves. In the first place, one or more of the non-test variables may be used as auxiliary cutting variables by means of which certain groups of individuals are completely excluded from the job in question. Thus, we may decide to employ no women in the job, no married men, no men who are not high-school graduates, or no men below 21 and over 45 years of age. In this case, absolute categorical limits are set on these variables, and at the specified point an absolute separation is made between a group which will be accepted and a group which will not. This type of procedure can usually be justified only on administrative rather than on scientific grounds. Even in such a dichotomy as sex, there is an overlapping of the groups in almost every measurable behavior function. In the quantitative variables such as age and education, the overlapping of adjacent groups is obviously very great. Using absolute rules which set critical scores on these variables will clearly exclude some individuals who would be acceptable by any index of aptitude or performance. The procedure may be politically expedient, but it is usually on grounds of expediency alone that it must be justified.

A second procedure may be to treat non-test variables as factors for which no absolute prescription is made, but as factors which are considered separately from the test scores. These variables become, then, essentially modifying factors which are taken into account in interpreting test results. They are treated in a clinical fashion in arriving at a judgment as to the administrative action. Thus, individuals over 40 years of age may be rejected unless they show exceptionally high aptitude. Or a specially high score on tests of intellectual functions may be required of the individual who is not a high-school graduate.

The procedures may be quite unspecified, or they may be formalized into a system of credits and allowances for the non-test factors. This approach is less formalized and rigid than the cutting score procedure. However, it also seems to lack any very good basis in statistical theory. It has a rule-of-thumb character which hardly recommends it.

The third possibility for the non-test variables is, of course, to incorporate them with the test variables into the score composite which is used to predict success on the job. If such a variable as age or education has been shown to predict success on the job, the validity data for that variable provide as good basis for weighting it in a regression equation as do the data on any test score. If it has not been shown to predict success on the job, there is no occasion to worry about that variable anyhow. In most cases, we may either expect the variable to bear a linear relationship to job success or else transform the measure of the variable into some derived score which does show such a linear relationship. Thus, if the optimum age for achieving success in a particular job is 28, the individual's deviation from that optimum may be used in the regression equation in place of his actual age. If necessary, a non-linear transformation may be used. If the non-test variable is qualitative in nature, it may be translated into a quantitative coding or rating. Thus, a description of the applicant's job experience may be rated by an interviewer or by an analyst of the application blank, and this rating can become the variable which enters into the composite score. In this way, the advantages of multiple regression techniques can be extended to other variables in addition to the actual test scores. Except where matters of administrative convenience suggest one of the alternative procedures, this way of handling auxiliary non-test data is probably to be preferred.

### THE FORM OF TEST SCORE TO BE REPORTED

One administrative decision which must be made in any testing program is the form of score to be used in reporting results. This problem comes up with reference to the scores on a single test; it also arises with regard to the composite scores which are derived from the complete battery and which represent predic-

tions of success in a particular job category. We shall consider these problems in turn.

In a selection program based on the system of weighted score composites, single test scores are used primarily (1) in the computation of the composite scores and (2) for research analyses. The problem is whether to use raw test scores for these purposes, or to transform the scores into some other form. The advantages and limitations of a single-digit score for record keeping and research have been discussed in Chapter 9. At that time some of the considerations for routine testing operations were also mentioned.

The final composite aptitude score for a particular job category is not a tool for obtaining some further type of score value, and its use in research is subsidiary to its use in administrative decisions about the individual. The considerations determining the type of score to be reported here are, therefore, primarily those relating to maximum administrative convenience and ease of interpretation. We must report the score which is most readily usable in the task of arriving at administrative decisions about each individual. Two characteristics appear desirable: The system of scores should be simple, and the system should be such that a given numerical score has the same meaning for different job categories.

The numerical value of the composite score derived from a weighted battery of tests is entirely arbitrary in any case. The numbers which result depend on the absolute size of the test weights. These can be adjusted by any multiplying factor and by the addition of any desired constant. It seems very desirable, therefore, that the constants be chosen in such a way that the mean score and the variability of scores are the same for each different composite score being computed from the data. A given number will then indicate a specified degree of excellence relative to the total array of applicants. The exact magnitude of the numbers to be used is a relatively minor matter, so long as the final grouping is fine enough to permit the necessary flexibility of adjustment of administrative decision. A single-digit score may often provide a somewhat coarse grouping, since raising or lowering the standard by a single point on such a scale may make a difference of as much as 20 per cent in the per-



centage of the group declared qualified. More than two digits (i.e., a possible range from 0 to 99) probably represents an unwarranted refinement and an assumption of greater accuracy than actually exists in most cases.

An alternative to the type of standard score discussed above may appeal to some. This is the percentile rank in terms of some defined group, usually the general population of job applicants. The concept of "percentage of individuals in the group whom he excelled" is a suitably simple and straightforward one. However, the steps in this scale of numerical values do not correspond even approximately to equal increments of ability. This type of score is therefore very difficult to use in any program of personnel classification. In a classification program, we are interested in differences in aptitude for different jobs, and it is difficult to evaluate and compare these differences directly from percentile scores.

The personnel psychologist will have to decide whether the standard scores which he uses are to be normalized or merely equated in terms of mean and standard deviation. This is probably not a very critical decision, nor will it affect his program greatly one way or the other. If data are available for a whole group of applicants, the assumption of normality appears fairly reasonable. The use of a set of normalized scores guarantees the comparability of meaning of a particular score for all job categories throughout the whole range of score values, rather than equating merely in terms of mean and standard deviation. The use of a set of converted scores which has a normal distribution also has some advantages for further statistical research, since the assumption of normality is made in the development of certain formulas.

Once a system of converted scores has been adopted, a good deal of educational activity is often required and may well be expended in getting over the meaning of different numerical score values to the administrators who use them or who control decisions as to the continuation of the selection program. When this education begins to be effective, there should be a good deal of reluctance to change the score system. In a sense, the standard scoring system tends to become one of the assets of the



psychologist, representing part of his "good will," and it should not lightly be abandoned.

In general conclusion, a one- or two-digit system of normalized standard scores appears to provide a simple, efficient, and uniform pattern for reporting score composites. The choice of number of digits and simultaneously of fineness of grouping should be made in terms of the need for flexibility and close adjustment of assignment percentages in the groups being worked with. Once a system of scores has been adopted and the administrative group has been trained to understand and use it, the system should be changed only for very good reason.

### FUNCTION OF THE PERSONNEL PSYCHOLOGIST IN THE ADMINISTRATIVE USE OF TEST SCORES

Finally, we must ask: What role should the personnel psychologist play in the administrative use of test scores? Does his function cease when he has determined an aptitude score for individual X for one or several job specialties? Should he in addition provide some interpretation of X's scores and some recommendation as to what action should be taken with regard to X? Should he be given responsibility for the final administrative decision with regard to X's job assignment?

For the psychologist merely to provide scores for each individual hardly seems adequate. Certainly, he has an additional responsibility for providing materials and training which will help the actual administrator in his use and interpretation of these scores. Furthermore, the psychologist should be a participant in any discussion of the way in which the scores are to be used and should participate in formulating the "ground rules" by which the scores are to be used. His voice should be heard when the administrative procedures are being set up covering such points as (1) multiple regression versus multiple cut-off procedures, (2) daily quota versus predicted yield plan for balancing supply and demand, (3) extent to which factors other than the psychologist's assessment of aptitude are to be considered. Once these points have been agreed on, the administrative decisions on personnel assignment may well be handled by personnel not directly concerned with the testing program.

This is acceptable particularly when the decisions are simple ones, such as those involved in a straight selection program, and when the person in charge of administrative disposition of each case is informed about and in sympathy with the testing program.

At the opposite extreme, the psychologist may be given final responsibility for the administrative decision on the placement of each man. It must be recognized, however, that assumption of such responsibility by the psychologist implies that he must become informed about many aspects of the operation of the organization which bear little or no relationship to psychological techniques for personnel selection. In particular, he must be well acquainted with the problems of personnel supply and demand in the organization. He must concern himself with quotas, flow figures, and a host of administrative details. He must be informed of anticipated changes in amount or type of activities within the organization, and he must be able to anticipate projected needs. There is no reason why he cannot do these things. However, if the staff for running a personnel selection program is limited, it may often seem desirable to limit the personnel psychologist to technical aspects of aptitude measurement and leave the administrative use of the test results to others.

Perhaps the appropriate utilization of the personnel psychologist should be outlined as follows:

1. He should devise, improve, and do research on selection and classification procedures.
2. He should plan for and supervise the administration of aptitude tests and the extraction of aptitude scores from these.
3. He should participate in determining policy and administrative routines for the utilization of test results.
4. He should provide test results in a form which will, as far as possible, suggest or specify the final administrative action.

## *The Personnel Selection Program and the Public*

A basic fact that every personnel psychologist needs to appreciate, whether he is working in industry, in civil service, or in the armed forces, is that the broad administrative decisions which determine the conditions under which he is to work and even the question of whether he is to continue to work will be made not by him but by his administrative superiors. Some person or persons in the top levels of management will have the power to decide that there is to be a personnel selection program and that psychological tests are to be used. Someone will obtain funds for the program. Someone will decide how much time of job applicants is to be made available for testing, and how much testing can be done with employees already on the job. The psychologist may be consulted on all these points. He will certainly have an opportunity to express his opinions and offer his recommendations. But the general decisions on matters of policy will not be his.

This dependence on others for the continuing support of his activities raises a new set of practical problems which the personnel psychologist must face. They are problems of salesmanship. A personnel program necessarily includes a selling program, to guarantee continuing acceptance of and support for the program. It is of critical importance to sell the program to those members of top management who have powers of life or death over the program. It is only slightly less important to sell the program to all those whose cooperation is involved in some degree in the operation of day-to-day testing and in the carrying out of research projects. Plant managers or post commanding officers who are required to supply space and equipment for testing, supervisors or instructors who are asked to provide cri-

terion ratings, experts who are asked to give time to the preparation or review of test materials, workers on the job who are called in to take research tests, all should feel that the program is worth while and that it will contribute to the well-being of themselves and their organization.

The present emphasis on selling is not intended to minimize the importance of what is sold. No defense is being made for skillful salesmanship applied to a shoddy, inferior product or service. In personnel selection, as in most fields, there is no lack of polished individuals who present in a compelling manner some completely unscientific and unvalidated technique. Quacks specializing in graphology, phrenology, physiognomy, or divine intuition are found in abundance. It is often true, unfortunately, that the best salesmanship is applied to the poorest product. The temperament which is disposed to careful and exacting research tends not to take kindly to or have a gift for promotion. But it is just this sound scientific worker who must in self-defense develop effective promotion for his service. The layman does not have the background to discriminate between effective personnel research and quackery. He must be educated and trained to discriminate between the tested results of a sound personnel system and the unfounded claims of the quack. The more scientific and rigorous a personnel research worker is, the more important it is for him carefully to consider the public relations side of his work.

The effective public presentation of the research basis for a personnel selection program is not an easy task. The research statistics involved in test validation and multiple prediction are not directly useful for this purpose. Furthermore, many of the direct indices of practical effectiveness of a program are confused by other factors. We must consider at this time the problems in effectively presenting the values of a testing program.

The first point to realize is that the individuals to whom a personnel program is to be sold are not statisticians. They usually have neither the training nor the inclination to delve into the subtleties of statistical procedures or statistical results. The standard statistics of test validation and multiple prediction are almost worse than useless in any promotion of a testing program. Correlation coefficients almost always produce in the

layman either misunderstanding or bafflement. In neither case is a desirable attitude toward the testing program achieved. The fact that they are not understood often arouses a defensive reaction of resentment in the audience, and a feeling that something is being "put over on" them. They are likely to confirm the impression that the psychologist is a "long-hair" who is off in a mystic world apart from reality and who does not understand practical affairs and the realities of a business enterprise. In his relations to his lay superiors and colleagues, the personnel psychologist may be well advised never to exhibit a correlation coefficient or mention a regression weight. The statistics to be used in public relations and promotion are almost entirely distinct from the statistics of personnel research.

How shall we present the results of a testing program to the lay audience? What statistical procedures or indices will be effective for this purpose? The indices must be simple. They must be readily understood. They should preferably lend themselves to simple visual presentation.

### SHOWING THE RELATIONSHIP OF TEST SCORES TO JOB SUCCESS

The basic source from which almost any material for representing the success of a personnel selection program must be derived is the bivariate frequency distribution showing the relation between some test or score composite on the one hand and some criterion variable on the other. In the more general case, in which the criterion is expressed as a continuous variable, this takes the form shown in Table I. This is the basic tabulation from which a measure of correlation is computed. When the criterion variable is a dichotomy, such as graduation versus elimination in training, or when it seems desirable to reduce the criterion score to a dichotomy for purposes of presenting the results, the data take the form shown in Table II. Table II was constructed from Table I by specifying that a criterion score of 30 or above represented success in the job, whereas a score below 30 represented failure. Obviously, Tables I and II as they stand do not represent satisfactory materials for use in promotional activities. We must consider what information can be abstracted from them for this purpose.



TABLE I. RELATIONSHIP OF APTITUDE TEST SCORE TO TRAINING SCHOOL GRADES

Aptitude Test Score	Grades in Training School (Criterion Score)										Total
	12-14	15-17	18-20	21-23	24-26	27-29	30-32	33-35	36-38	39-41	
85-89									1	1	2
80-84								2	4	6	12
75-79						5	8	15	10	5	43
70-74					4	28	28	15	12	3	90
65-69				2	14	62	36	19	12	2	147
60-64				10	28	79	55	49	8	1	230
55-59			3	17	52	68	37	22	6		205
50-54		3	10	20	43	25	15	3	1		120
45-49	1	5	25	32	18	3	1				85
40-44	1	12	15	12	3						43
35-39	3	6	8	1							18
30-34	3	1	1								5
Total	8	27	62	94	102	270	180	125	54	18	1000

TABLE II. RELATION OF APTITUDE TEST SCORE TO SUCCESS OR FAILURE IN TRAINING SCHOOL

Aptitude Test Score	Training School Record		
	Number of Successes	Number of Failures	Total Number
85-89	2		2
80-84	12		12
75-79	38	5	43
70-74	58	32	90
65-69	69	78	147
60-64	113	117	230
55-59	65	140	205
50-54	19	101	120
45-49	1	84	85
40-44		43	43
35-39		18	18
30-34		5	5
Total	377	623	1000

Corresponding to each numerical statement of relationship, there are usually one or more graphic forms for representing those numerical facts. The graphic forms of representation are of particular importance in a promotional program. The visual representation of facts, if well done, is ordinarily much more

effective than the representation of those same facts in numerical or tabular form. Graphs and pictograms are almost universally familiar and make a general appeal. If it is well conceived and kept at a sufficiently simple level, the spatial representation makes a most important supplement to the numerical. In the discussion which follows, therefore, possible forms of graphic representation will be discussed along with the numerical indices which they represent. Some of the problems and pitfalls of graphic representation will also be considered.

Working from Table I and retaining the continuous distribution of criterion scores, we find two ways in which the facts of relationship may be exhibited to a lay audience. The first procedure is really no more than a simplification of the original bivariate distribution by reducing the number of categories in criterion score, test score, or both. Instead of ten to twenty score categories for each variable, we may reduce the number to four or five. At the same time, in order to facilitate comparisons from one test to another, the categories are usually defined in terms of the proportion of the group in each. Thus, we may prepare a table, dividing both test and criterion scores into highest quarter, second quarter, third quarter, and lowest quarter. Taking those in any quarter on test score, we can show what percentages of them are in the highest, second, third, and lowest quarters, respectively, on criterion score. Table III shows the data of Table I consolidated into quarters on both test and criterion

TABLE III. RELATION OF APTITUDE TEST SCORE TO TRAINING-SCHOOL GRADES  
(Analyzed in terms of quarters of the group for both test score and grades.)

<i>Quarter of Group in Test Score</i>	<i>Quarter of Group in Grades</i>			
	<i>Bottom Quarter</i>	<i>Third Quarter</i>	<i>Second Quarter</i>	<i>Top Quarter</i>
Top quarter	6	51	78	115
Second quarter	21	67	83	79
Third quarter	49	83	69	49
Bottom quarter	174	49	20	7

score, and shows actual frequencies in each cell of the table. Table IV shows the frequencies in each row translated into

TABLE IV. RELATION OF APTITUDE TEST SCORE TO TRAINING-SCHOOL GRADES  
(Per cent of those in each quarter on test score falling into each quarter on training-school grades.)

Quarter of Group in Test Score	Per Cent Falling in Quarter in Grades			
	Bottom Quarter	Third Quarter	Second Quarter	Top Quarter
Top quarter	2.4	20.4	31.2	46.0
Second quarter	8.4	26.8	33.2	31.6
Third quarter	19.6	33.2	27.6	19.6
Bottom quarter	69.6	19.6	8.0	2.8

percentages of the total frequency in that row. Table IV is a form which might be used to represent the validity of a testing

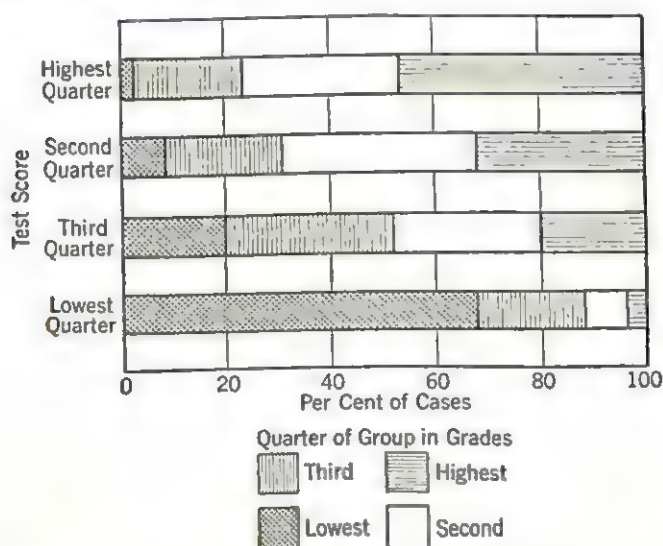


FIGURE 1. Relation of aptitude test score to training-school grades.

procedure. The data of Table IV are shown graphically in Figure 1.

One objection to the procedure of grouping data by fourths or fifths of the total group, from the point of view of the statistician, is that these fractions do not ordinarily correspond to equal steps of ability. Because most abilities yield a distribution of scores tending toward the normal distribution, the cases tend to pile up near the center of the distribution, and a given percentage of cases at one extreme covers a greater range of scores than does the same percentage near the middle of the distribution. It is possible to choose the percentages falling in the four or five test score categories so that they correspond approximately to equal distances on the abscissa of the normal frequency curve. Thus for five categories the percentages would be 7, 24, 38, 24, and 7, instead of 20, 20, 20, 20, and 20.

Equality of scale units is probably more of theoretical than of practical importance. The practical employer probably gets a more meaningful picture from equal percentages of his total group of applicants than he does from equal units on some hypothetical scale of ability. He may well be confused by the fact that the extreme groups contain small numbers and the middle groups large numbers. If the unequal groups *are* used, some additional educational activity will ordinarily be necessary to present the normal frequency distribution as characteristic of human abilities.

One promotional advantage of using percentages based on the normal curve is that the resulting groups are somewhat more widely spaced with regard to the trait measured by the test and consequently the discrimination between the different groups in criterion score is sharpened. Again, the more groups one uses, the more striking will be the difference in performance for the extreme groups.

One specific situation in which the use of unequal percentages in the several score categories is likely to appeal to the personnel psychologist is that in which he is carrying out all his analyses with single-digit scores. The advantages of using normalized single-digit scores for research analyses have been discussed in Chapter 9. If scores are already in single-digit form, these scores may often be used directly for tabular and graphic representation.

Table IV and Figure 1 show about the simplest representation which can be achieved when several categories are retained both for test score and for criterion. Most persons will probably agree that this representation is still somewhat involved for the lay reader. In particular, the comparison of several such tables or charts for different tests would not be a simple task. This complexity seems to be almost inherent in the material. It is one reason for devoting a good deal of the consideration in this chapter to dichotomized criterion variables.

One other procedure for representing the relationship of a continuous criterion variable to test score is to compute the average criterion score for those individuals in each test score category. The mean or median may be computed for each row of the bivariate frequency distribution, and these values may be plotted. The results for such a procedure are shown in Table V

TABLE V. MEDIAN TRAINING-SCHOOL GRADE FOR EACH LEVEL OF APTITUDE TEST SCORE

<i>Aptitude Test Score</i>	<i>Number of Cases</i>	<i>Median Academic Grade</i>
85-89	2	38.5
80-84	12	38.5
75-79	43	34.2
70-74	90	30.9
65-69	147	29.3
60-64	230	29.4
55-59	205	27.9
50-54	120	25.4
45-49	85	21.6
40-44	43	19.2
35-39	18	17.5
30-34	5	14.0

and Figure 2. However, this form of presentation can hardly be recommended to honest personnel psychologists. It is misleading in the extreme. The tabulating and plotting of means ignores entirely the variability within a given score group. It conveys the impression of an almost one-to-one relationship between test and criterion score. This impression may be corrected in some measure by also giving some measure of variability of the criterion scores for each test score category. However, the addition of a measure of variability, as shown in Table VI



and Figure 3, tends once more to complicate the representation. We may well question whether Figure 3 would convey much direct meaning to the non-specialist.

We turn now to a consideration of the forms of representation of relationship when the criterion is obtained as or is translated into a dichotomy. The basic tabular representation is given in Table II. Table II is readily converted into a table showing

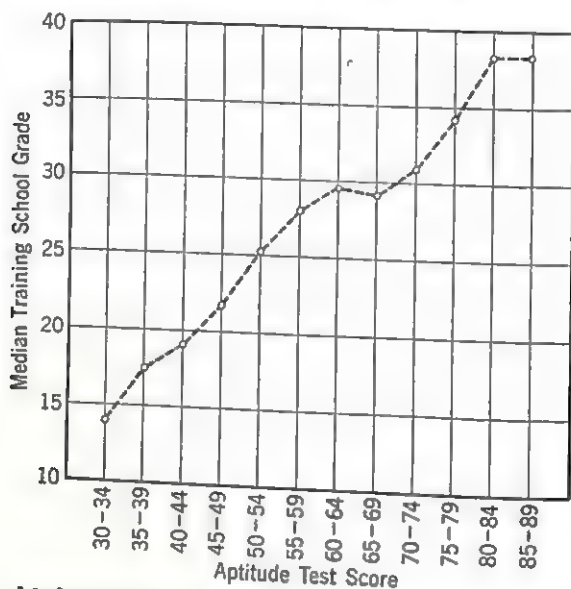


FIGURE 2. Median training-school grade at each level of aptitude score.

percentage failing for each test score category. This form of tabulation is shown in Table VII. It represents probably the simplest descriptive statement of the relationship between test and criterion which we have discussed so far. It lends itself readily to graphic representation as a bar chart, of the type shown in Figure 4. For purposes of comparing different tests and procedures, it may be desirable to establish a uniform procedure with regard to the percentage of cases to be included in each score category. The choice falls, once again, between including the same percentage of cases in each category (10, 20, or 25) or choosing percentages that will represent equal scale units on the abscissa of a normal curve. The considerations

TABLE VI. AVERAGE AND VARIABILITY OF TRAINING-SCHOOL GRADES FOR EACH LEVEL OF APTITUDE TEST SCORE

<i>Aptitude Test Score</i>	<i>Number of Cases</i>	<i>Median Academic Grade</i>	<i>Range of Grades, Middle 50% of Cases</i>
85-89	2	38.5	*
80-84	12	38.5	36.2-40.0
75-79	43	34.2	31.7-37.2
70-74	90	30.9	28.3-34.0
65-69	147	29.3	27.5-32.2
60-64	230	29.4	27.2-32.5
55-59	205	27.9	25.3-30.6
50-54	120	25.4	23.2-28.2
45-49	85	21.6	19.3-23.6
40-44	43	19.2	16.9-21.5
35-39	18	17.5	15.2-19.2
30-34	5	14.0	*

\* Too few cases to compute a meaningful measure of variability.

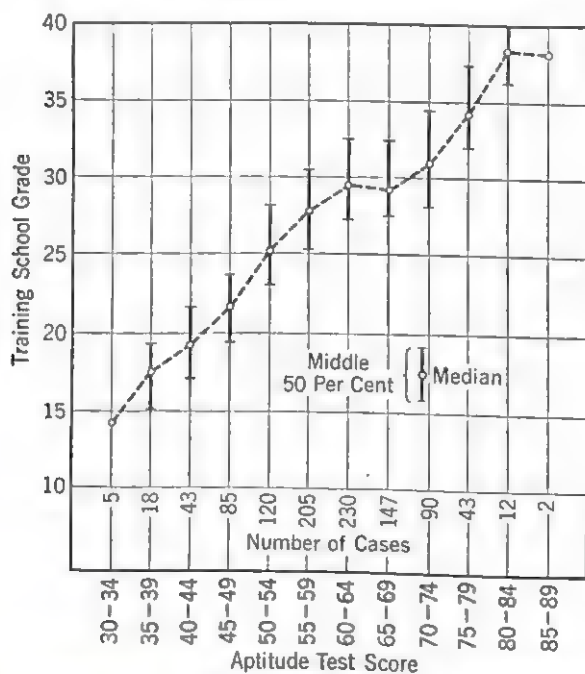


FIGURE 3. Relationship of aptitude test score to training-school grades. (Median grade and range of middle 50 per cent at each aptitude score level.)

TABLE VII. PER CENT OF APPLICANTS AT EACH APTITUDE SCORE LEVEL FAILING IN TRAINING SCHOOL

<i>Aptitude Score</i>	<i>Number of Cases</i>	<i>Number Failing</i>	<i>Per Cent Failing</i>
85-89	2		0
80-84	12		0
75-79	43	5	12
70-74	90	32	35
65-69	147	78	53
60-64	230	117	50
55-59	205	140	68
50-54	120	101	84
45-49	85	84	99
40-44	43	43	100
35-39	18	18	100
30-34	5	5	100
Total	1000	623	62.3

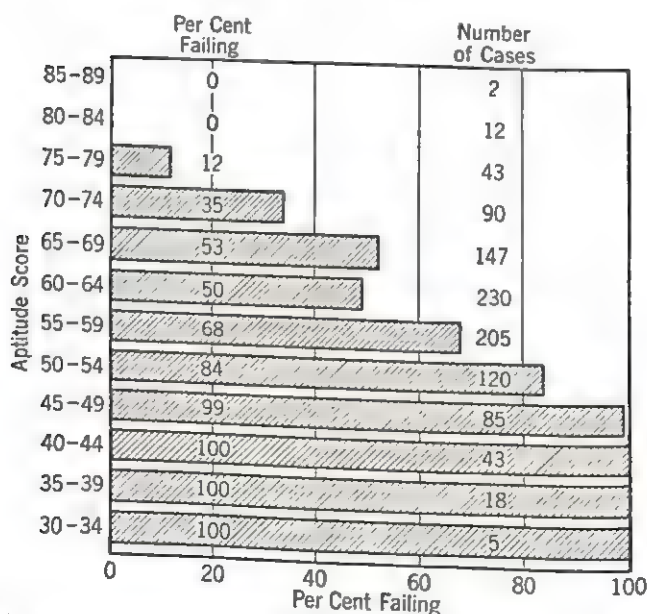


FIGURE 4. Percentage of applicants at each aptitude score level failing in training school.

entering into this choice have already been discussed in connection with the representation of continuous criterion scores.

The table of percentages succeeding at each score level and the bar chart which gives a pictorial representation of these facts provide a simple and satisfactory descriptive representation of the effectiveness of a test for selection purposes. The representation is not entirely immune to a certain amount of manipulation by the statistician. Where the dichotomy is an entirely arbitrary one, the choice of criterion score at which to make the division into two groups will have a good deal of influence on the effectiveness of the resulting picture. In the graphic representation, the decision whether to plot percentage succeeding or percentage failing and the choice of the scale in terms of which the percentages are plotted may have a good deal of influence on the way the resulting picture looks. However, the procedure is generally satisfactory.

### INDICES OF PRACTICAL EFFECTIVENESS OF SELECTION PROCEDURES

In addition to a simple graphic representation of the general trend of relationship, it is often desirable to present a simple index of the practical effectiveness of a particular testing procedure. The management may raise the question: If we employ this procedure, what decrease may we expect in the percentage of unsatisfactory persons employed? We require some statistic that will forecast the improvement to be expected if the selection technique is employed.

At this point, we must observe that any forecast as to what *will* happen as the result of a personnel selection program is a very hazardous undertaking. It necessarily involves forecasting the course of future events, on the assumption that other factors will remain the same as they have been in the past. Other factors have an unpleasant habit of not remaining the same, and the changes in these other factors may distort and even completely mask the effect of improved personnel selection procedures.

In the Army Air Forces one indicator of the practical success of the program for selecting pilots, navigators, and bombardiers

might have been the reduction in percentage of failures and eliminations from training. Knowing the elimination rate at each composite score level and the percentage of individuals having that composite score, it was possible to predict the percentage who would be eliminated in future classes if any specified minimum score were required in order to qualify. A set of tables could have been prepared to show the predicted failure rate for each cutting score. This prediction would necessarily have been based on the assumption that the elimination rate for those at a specified aptitude level would remain unchanged from class to class.

However, the elimination rate in pilot training, for example, showed wide cyclical changes unrelated to personnel selection procedures. Rates went up and down in wide swings and were related in an interesting way to certain speeches by high-ranking officers in the training program. When top command pounded on the table and insisted that every possible pilot was needed to fight the enemy, elimination rates went down. When, somewhat later when pilots were in abundant supply, the message was that standards must be maintained at all costs, the elimination rates went up again. The criterion measure was not a stable one, and its sensitivity to influences of the type we have mentioned made predictions of future elimination rate a very shaky foundation for promoting the personnel selection program.

In almost any personnel situation, factors other than the quality of entering personnel will affect a criterion measure to some extent, and these outside factors are likely to blur direct beneficial effects from the personnel procedures. They may either exaggerate or destroy the gains. Changes in the general level of business activity would distort any prediction of volume of sales in a selected group of salesmen. Changes in employment opportunities would distort a criterion measure expressed as rate of employee turnover. The introduction of a new union organization and union rules may upset predictions of employee output. Obsolescence or replacement of machines may completely upset predictions of production or spoilage. Predictions based on the single factor of personnel selection are shaky things, and the personnel psychologist is probably wise not to rest his case too



dogmatically on them. However, we shall need to consider several such indices.

One index of the practical effectiveness of a selection procedure is the comparison of the percentage achieving success on the job in a selected group and the percentage achieving success in the unselected group of applicants. These percentages can be obtained in either of two ways. They may be computed directly from the table giving frequency of success or failure at each score level (Table II). Or, on the assumption that the bivariate distribution of test score versus criterion measure is normal, they may be obtained from a set of prepared tables.<sup>1</sup> The tables require knowledge of the correlation between predictor and criterion, the percentage to be excluded on the basis of the selection device, and the percentage of job failures in the unselected group of applicants. The tables provide a more expeditious way of calculating the desired percentages of job success, and they have the special advantage that they can be applied when the complete table of test versus criterion scores is not available. The values which the tables provide are theoretical ones, whereas calculation from the actual bivariate frequency distribution preserves both the tangibility and the irregularities of the original score data.

The type of information given by this type of analysis is shown in Table VIII and Figure 5. These are based upon the actual frequencies in Table II. The essential facts which are related here are the percentage of applicants who would have been rejected and the improvement which would have been effected (other factors remaining unchanged) in the job failure rate. The table can be read as follows: If we had set a minimum score of 40, which would have disqualified 2.3 per cent of applicants, the percentage of job failures would have been reduced from 62.3 to 61.4; if we had set a minimum score of 50, which would have disqualified 15.1 per cent of applicants, the percentage of failures would have been reduced from 62.3 to 55.7, etc. Knowing what the practical situation will bear in the way of elimination of applicants, it is possible to estimate the resulting improvement

<sup>1</sup> H. C. Taylor and J. R. Russell, "The Relationship of Validity Coefficients to Practical Effectiveness of Tests in Selection," *J. Appl. Psychol.*, **23**, 565-578 (1939).

in job success. Though the table and graph may be a little complex, the result is expressed in a form which has quite direct

TABLE VIII. EFFECT UPON TRAINING-SCHOOL SUCCESS OF USING APTITUDE TEST SCORE AS QUALIFYING SCREEN

(Rejection of applicants and reduction in training-school eliminations resulting from different qualifying scores, as derived from sample of 1000 cases.)

<i>Minimum Aptitude Test Score to Qualify</i>	<i>Per Cent Disqualified by Test</i>	<i>Predicted Elimination during Training School</i>
85	99.8	0.0
80	98.6	0.0
75	94.3	8.8
70	85.3	25.2
65	70.6	39.1
60	47.6	44.3
55	27.1	51.0
50	15.1	55.7
45	6.6	59.7
40	2.3	61.4
35	0.5	62.1
30	0.0	62.3

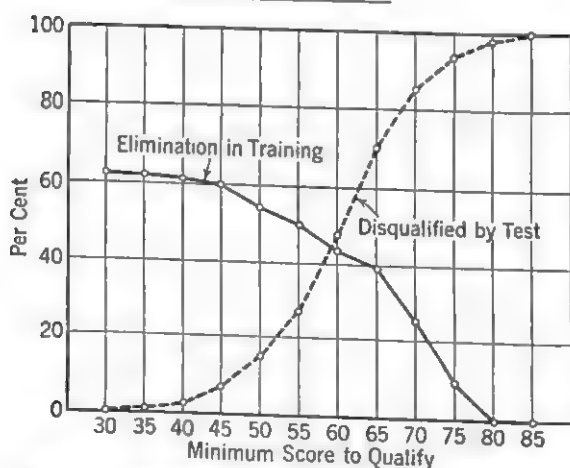


FIGURE 5. Effect of minimum qualifying score upon elimination rate and disqualification rate.

practical meaning. The only catch lies in the personnel psychologist staking his reputation that the future *will bear out* the change indicated by past data.

A somewhat different indication may be obtained from a table of the cumulative percentage of job failures and of job successes. The percentage of future job failures who are eliminated by a cutting score set at a specific point may be spoken of as the *utility* of the selection device, and the percentage of future job successes eliminated may be spoken of as the *cost*.<sup>2</sup> The greater the utility for a given cost, the better the test. A cost-utility table of the data from Table II is given in Table IX, and the

TABLE IX. EFFECTIVENESS OF SELECTION BY APTITUDE TEST

(Cost of selection program, in terms of potential successes rejected, compared with utility, in terms of potential failures rejected for different qualifying standards. Sample of 1000 cases.)

<i>Minimum Aptitude Test Score to Qualify</i>	<i>Cost (Per cent successes rejected)</i>	<i>Utility (Per cent failures rejected)</i>
85	99.4	100.0
80	96.3	100.0
75	86.2	99.0
70	70.8	93.9
65	55.7	81.5
60	22.6	62.8
55	5.6	40.3
50	0.3	24.1
45	0.0	16.6
40	0.0	3.7
35	0.0	0.8
None	0.0	0.0

facts are presented graphically in Figure 6. Cost-utility provides a means of making direct comparisons of different tests and selection procedures, by establishing a fixed cost value and determining the utility for each procedure. Thus, if we decide that a wastage of 20 per cent of potential successes is a reasonable figure, we can determine the percentage of potential failures which can be screened out at that cost in potential successes. Obviously, the higher this percentage, the more satisfactory the selection procedure is. The concepts of cost and utility can be

<sup>2</sup> The author was introduced to the concepts of *cost* and *utility* by Dr. Joseph Berkson, then Colonel, M. C. and Chief, Statistical Division, Office of the Air Surgeon.

interpreted in practical terms much more readily than can correlation coefficients. Though cost-utility appears to have no value as a statistic for test analysis, it may be a usable conceptual framework for public relations.

Where the criterion which is being predicted is the successful completion of some training course, another form of presentation which is sometimes effective shows the number who must be

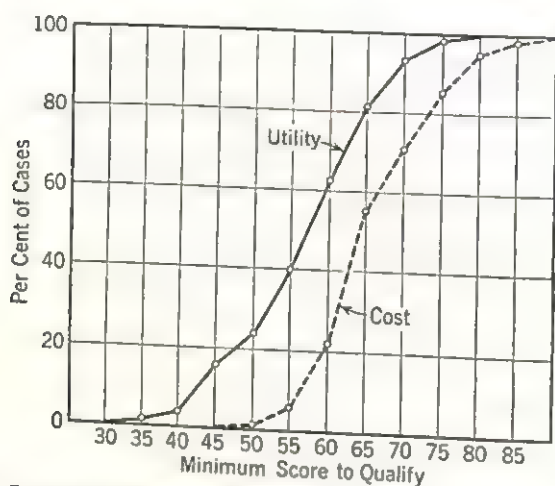


FIGURE 6. Cost-utility index of effectiveness of aptitude test. (Potential successes excluded versus future failures screened out by test.)

entered into training in order to produce 100 who successfully complete the course. This number is shown for the unselected group of applicants, and for groups screened by different minimum values of aptitude test score. As screening becomes more selective and as the anticipated failure rate consequently drops, the number that must be entered into training drops. As a useful supplement to these facts, one can show the total number who would have to be tested to yield the 100 graduates. This can be determined by dividing the number to be entered into training by the percentage who can be expected to achieve the qualifying score. As the qualifying score is raised, the number to be tested will, of course, go up. The data of Table VIII have been analyzed in this way, and the results are shown in Table X and Figure 7. An examination of these two sets of facts, together

TABLE X. EFFECT OF APTITUDE REQUIREMENTS. NUMBER OF ENTERING STUDENTS AND NUMBER OF APPLICANTS REQUIRED IN ORDER TO YIELD 100 SUCCESSFUL GRADUATES

(Based on sample of 1000 trainees.)

<i>Minimum Aptitude Test Score to Qualify</i>	<i>Number Entering Training to Yield 100 Successful Graduates</i>	<i>Number of Applicants Tested to Yield 100 Successful Graduates</i>
80	100	7100
75	110	1930
70	134	910
65	164	500
60	180	340
55	204	280
50	226	266
45	248	265
40	259	265
35	264	265
30	265	265

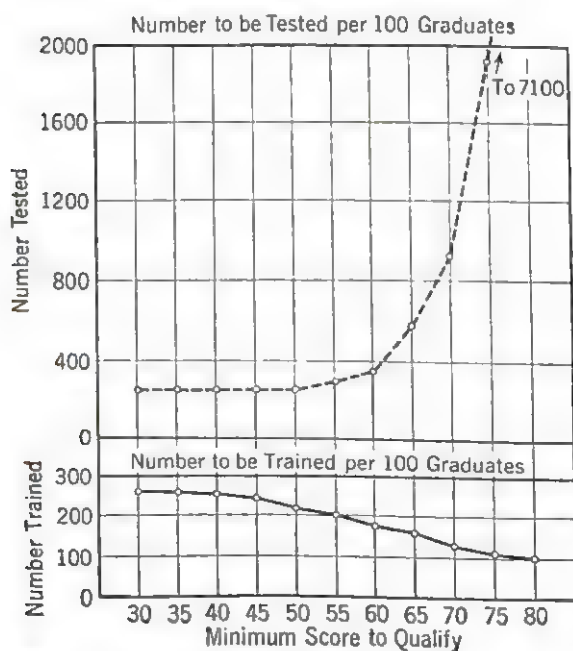


FIGURE 7. Effect of aptitude test requirement on numbers to be tested and numbers to be trained.



with background information on the supply of applicants, the cost of testing each applicant, and the cost of training each man, provides a sound basis for deciding what degree of selectivity is most advantageous in a particular case.

### PUTTING THE FACTS TOGETHER

Organizing the validation data in terms of understandable statistics instead of research statistics is only one step in presenting the personnel selection program to the public. It is then important to translate the results into pictorial terms, and to correlate the isolated facts into a simple but compelling story. Some suggestions for graphic representation of the validity data are given by the figures in this chapter.<sup>3</sup> These figures are somewhat limited by the black and white and the limited size of the printed page. In a presentation being prepared for the eye of top management graphs should be of generous size and the dramatic effect enhanced by liberal use of color.

Though isolated tables and charts have their place, particularly as material to illustrate points made in discussions by the personnel psychologist, a further need is for public relations material which tells a complete story by itself. A self-contained "package" or "presentation" should be prepared which is sufficiently attractive to catch and hold the attention and sufficiently simple and straightforward for the lay reader. Visual attractiveness and ease of understanding are characteristics which go together in reading material, and factors which favor the one will usually favor or at least not interfere with the other. Among the factors which contribute attractiveness, we may mention the following:

1. A neat and professional-looking format. Good lettering, well-balanced pages. A generally "finished" appearance.
2. Liberal use of pictures, pictographs, and cartoons.
3. Reading matter in small doses, with pictorial matter breaking up the reading.

<sup>3</sup> For other types of graphic representation the personnel psychologist can refer with profit to such sources as W. C. Brinton, *Graphic Presentation*, Brinton Associates, New York, 1939.

4. Simple and colorful charts and graphs.
5. Reduction of non-essential details to a minimum.

Many of these factors also make for ease of comprehension. We can supplement them by mentioning the following:

1. Use of simple language and short sentences.
2. Elimination of technical statistical or testing jargon.
3. Careful organization of the presentation, so as to develop a single main theme from its introduction to its conclusion.

As soon as validation data begin to accumulate for a personnel selection program, they should be organized so that they may be presented to the interested public. When the data reach a satisfactory volume, when some special major project of validation has been completed, or when some important administrative decision is in the offing, a special effort can well be made to assemble the most complete and current information on the selection program for effective presentation to those whose support is needed for the continued success of the program.

### PROMOTION OF NEW PERSONNEL PROJECTS

Promotion of a proposed new direction of work or investigation presents a somewhat different problem from eliciting support for an existing program. With an existing program, the important thing is to exhibit the effectiveness of the instruments and procedures being used, and the gains which have resulted from using them. When a new project is being promoted, this type of direct evidence of the effectiveness of the psychologist's functioning is lacking. However, evidence of success in dealing with past problems is probably the strongest argument for support of new ones. In presenting a proposal for attacking a new problem, three aspects are often involved: (1) showing the existence and nature of the problem, (2) showing that it is a suitable one for the psychologist to be concerned with, and (3) showing that the psychologist has a good chance of success in dealing with it.

Problems dealing with personnel efficiency are sometimes recognized by management; sometimes they are not. Because of his special interest in the human factor, the psychologist should

be particularly sensitive to situations in which efficiency is below par because of poor employee selection, poor training procedures, poor morale and employee-supervisor relations, and equipment poorly designed from the point of view of the human factor. However, sensitivity by itself is not enough to discover and develop an understanding of these problems. A wide range of contacts with all phases of the organization in which he is working is also required. Part of this acquaintance is a natural outcome of any extensive program of job analysis, during which the psychologist should become well acquainted with all sides of the job which he is studying. Part of it grows out of effective contact and liaison with many persons at many levels in the organization. As opportunity permits, the personnel psychologist should devote available time to improving his knowledge of operations and conditions in his organization, with a view to identifying problems involving the human factor, to which his training might profitably be applied.

When the personnel psychologist has located a problem on which he believes he could make a valuable contribution, he then faces the necessity of documenting his belief. To do this, he must first organize the facts about the existing problem. He must gather evidence to show that there is a problem and that the present situation is unsatisfactory. Opinions, facts, and figures on such things as wastage, employee turnover and the like should be organized concisely but effectively, with supporting figures and tables.

Showing that the problem is one on which the psychologist might suitably work and with some prospects of success involves a demonstration of the human factor in the problem, and of previous success by psychologists in dealing with that type of human problem. Evidence on these points may come in part from the particular new problem. In large part, it will come from the previous work of personnel psychologists in that organization and in other organizations. Previous success in the same organization may be the most effective evidence. However, when work of a novel type is proposed which has not previously been carried out in that organization, the experience of other groups may be cited to show the probable fruitfulness of the enterprise.

## THE PERSONNEL PSYCHOLOGIST AND THE EMPLOYEE

The public relations activities discussed so far are those that are directed primarily to individuals in positions of authority over the personnel program. An effective promotional presentation can be very useful also in eliciting the cooperation of division managers, supervisors, and various specialists whose cooperation is needed for implementing the work of the psychologist. Understanding of the psychologist's goal and of his progress toward that goal are good bases for help and cooperation. However, there is a further public relations program for the psychologist in connection with the personnel of the organization and particularly with the employees and job applicants whom he must test. This is a problem of employee morale with specific reference to the testing program.

The personnel psychologist is almost inevitably a representative of management rather than of the worker. Even if a psychologist were employed by a labor union to work on selection problems, one suspects that the individual working man would think of him as working for "them" rather than for "us." Under these circumstances, the problem of eliciting employee cooperation for research testing, collection of criterion ratings, and the like is a very real one. Cooperation and morale do not grow overnight. They develop gradually over the months and years. They are the outcome of efforts by the psychologist to explain his functions on the one hand and to deal forthrightly with employees on the other. In his contacts with employees, the psychologist should endeavor to explain what he is doing as completely as possible within the limits of available time, existing audience background, and the needs for safeguarding his materials and procedures. He should give a simple picture of his work and admit his audience to participation in and understanding of it.

Most important for long-time morale, the psychologist should act with complete integrity in relation to employees. If he promises that certain results are to be kept anonymous, he should see that they are so. If he states that certain tests will be used only for research purposes, he should not permit any

other use to be made of them. If he reports that testing of present employees is "for the purpose of picking certain men for promotion and for special assignments," he should resist bitterly any effort by management to use those results as a basis for firing inept employees. He must not permit management to make him or his testing program "take the rap" for unpleasant managerial actions. In other words, he must be a person of integrity in his dealings with employees, so that his reputation will be that of a man who is to be trusted and a man whose actions correspond to his words.



## APPENDIX A

### *Solution of Normal Equations in Order to Determine Regression Weights and Multiple Correlation*

In using a set of tests to predict a criterion variable, the goal is to predict the criterion as accurately as possible. Maximum accuracy is defined in terms of least squares. The requirement is that the sum of the squared deviations of predicted from actual scores shall be a minimum, i.e.,

$$\Sigma(\hat{y} - y)^2 = \text{minimum}$$

It has been shown in Chapter 7 that this goal is achieved when the following set of equations is satisfied:

$$\beta_1 + \beta_2 r_{12} + \beta_3 r_{13} + \beta_4 r_{14} = r_{1y}$$

$$\beta_1 r_{12} + \beta_2 + \beta_3 r_{23} + \beta_4 r_{24} = r_{2y}$$

$$\beta_1 r_{13} + \beta_2 r_{23} + \beta_3 + \beta_4 r_{34} = r_{3y}$$

$$\beta_1 r_{14} + \beta_2 r_{24} + \beta_3 r_{34} + \beta_4 = r_{4y}$$

These are known as the set of *normal equations*. It can be seen that the set consists of  $k$  symmetrical linear equations, where  $k$  is the number of predictor variables. The  $\beta$ 's are the weights to be applied to the separate predictor variables, when these are expressed in the form of standard scores. The problem, then, is to solve the set of  $k$  equations for the  $k$  unknowns.

Two methods of solving these equations will be presented. The first is a compact version of the Doolittle procedure for solving normal equations. The solution in its present form is a rearrangement, in the interest of a more mechanical computing routine, of the abbreviated Doolittle method described by Dwyer.<sup>1</sup> The second is an iterative procedure based on the work

<sup>1</sup> Paul S. Dwyer, "The Solution of Simultaneous Equations," *Psychometrika*, 6, 101-129 (1941).

of Kelley and Salisbury.<sup>2</sup> The iterative procedure greatly reduces computational labor for problems involving a large number of variables, if the final weights need not be accurate to more than two decimal places.

Each of these methods will be illustrated by the same four-variable problem. The problem has been kept to a small number of variables in order not to confuse the exposition with too much computational detail. The development which follows presents *computational routines* and makes no pretense of providing an understanding of the mathematical bases of the methods.

The illustrative problem consists of data in which four tests have been used as predictors of a criterion measure of language achievement. The variables are:

Test 1	Verbal Reasoning
Test 2	Arithmetical Ability
Test 3	Figure Analogies
Test 4	Spatial Reasoning

The intercorrelations and validities are as follows:

	<i>Test</i>				
	1	2	3	4	Criterion
1	...	.58	.56	.52	.62
2	.58	...	.54	.47	.50
3	.56	.54	...	.56	.44
4	.52	.47	.56	...	.36

### METHOD 1—ABBREVIATED DOOLITTLE

This method, like the standard Doolittle solution and many other procedures for solving normal equations, operates by eliminating variables from the set of equations one after another until only one remains. The weight for the remaining variable is then given directly. The solution is then reversed to determine weights for each of the variables. The distinctive feature of the procedure described here is that it makes full use of the modern computing machine, which will carry out computations of the type  $(a - bc - de)$ . Many of the sequences of computa-

<sup>2</sup> T. L. Kelley and F. S. Salisbury, "An Iteration Method for Determining Multiple Correlation Constants," *Jour. Amer. Stat. Assn.*, 21, 282 ff (1926).

tion can be carried out on the computing machine, with no recording or copying, so that the labor and chance for error in transcribing are reduced to a minimum.

The solution of the four-variable problem given above is presented in full in Table 1. This table can be analyzed into

TABLE 1. ABBREVIATED DOOLITTLE SOLUTION

	1	2	3	4	5 Criterion	6 Check
IA	1.0000	.5800	.5600	.5200	.6200	3.2800
IIA		1.0000	.5400	.4700	.5000	3.0900
B		.6636	.2152	.1684	.1404	1.1876
C		1.0000	.3243	.2538	.2116	1.7896
IIIA			1.0000	.5600	.4400	3.1000
B			.6166	.2142	.0473	.8781
C			1.0000	.3474	.0767	1.4241
IVA				1.0000	.3600	2.9100
B				.6124	-.0145	.5980
C				1.0000	-.0238	.9765
D				1.0000	.0850	
E				1.0000	.1900	
F				1.0000	.4746	

three parts. Line IA is made up of the first row from the original table of intercorrelations and validity coefficients, with the addition of a check sum. Sections II, III, and IV are the successive stages in reducing the number of variables to three, two, and finally one. Each of these sections involves the same sequences of operations, which will presently be followed through in more

detail. The final section consists of lines D, E, and F, which carry through the back solution for determining the actual regression weights. Step-by-step instructions will now be given for carrying through the solution.

1. Prepare a computing sheet. If  $k$  is the number of predictor variables, the computing sheet will require  $4k - 3$  rows and  $k + 2$  columns. Arrange the sheet as shown, with an initial section of *one* row,  $k - 1$  sections each of three rows, and columns as indicated in Table 1. If the number of predictors is increased this will require, for each added predictor variable, (a) one added column, (b) one added section of three rows, A, B, and C, and (c) one row added to the final section giving the back solution.

2. Enter in row IA the values from the first row of the original table of intercorrelations and validities. For the diagonal term, which goes in IA1, enter unity. In the final check column enter the sum of all the other entries in row IA.

3. In row IIA enter the values from the second row of the table of intercorrelations and validities. Start with the diagonal term, which goes in position IIA2, and enter only the terms to the right of the diagonal. The entry in the final check column is the sum of the entries in row IIA and the entry in IA2.

4. Fill in all the following A rows in the same way, starting in each case with the diagonal term, which will always be unity. The check sum will be in each case the sum of the terms in the row plus all A entries in the column above the diagonal term.

5. To find the entry at IIB2, subtract from IIA2 the square of IA2. That is  $IIB2 = IIA2 - (IA2)^2$ . The other terms in row IIB are  $IIB3 = IIA3 - (IA2)(IA3)$ ,  $IIB4 = IIA4 - (IA2)(IA4)$ , etc. This process of multiplication and subtraction is carried out, preferably carrying out the whole operation on the computing machine, for each of the entries in row IIB.

6. To find the entries in row IIC, each entry in row IIB is divided by IIB2. Each entry in row IIC is derived from the entry in the corresponding column of row IIB.

7. To check the accuracy of calculations in either of the above two steps, the sum of all entries excepting that in the check column is obtained. This sum should equal the value in the

check column, except for rounding errors. Enough decimal places must be retained in calculations so that rounding errors will not affect the final result.

8. To find the entry at IIB3, subtract from IIIA3 both  $(IA3)^2$  and  $(IIB3)(IIC3)$ ; thus,  $IIB3 = IIIA3 - (IA3)^2 - (IIB3)(IIC3)$ . Similarly  $IIB4 = IIIA4 - (IA3)(IA4) - (IIB3)(IIC4)$ . The entry in the criterion column may be called IIB5. It is obtained by the set of operations  $IIIA5 - (IA3)(IA5) - (IIB3)(IIC5)$ . Again, each complete sequence of multiplications should be carried out on the computing machine.

9. The entries in row IIC are again found from those in IIB by dividing each term in IIB by IIB3. The check step follows as in (7) above.

10. The set of operations outlined in (5) and (6) and again in (8) and (9) is repeated a total of  $k - 1$  times. With each repetition, the number of products to be subtracted from the entries in the A row is increased by one.

11. Finally, a C row is obtained which has only three entries, one entry of 1.0000 in the body of the table, an entry in the criterion column, and one in the check column. The entry in the criterion column is the regression weight for the last variable, in this case  $\beta_4$ . That is  $\beta_4 = IVC5$ .

12. The values for  $\beta_3$ ,  $\beta_2$ , and  $\beta_1$ , given in rows D, E, and F, respectively, are obtained from the following expressions:

$$\beta_3 = IIC5 - (IIC4)(IVC5) = D5$$

$$\beta_2 = IIC5 - (IIC3)(D5) - (IIC4)(IVC5) = E5$$

$$\beta_1 = IC5 - (IC2)(E5) - (IC3)(D5) - (IC4)(IVC5) = F5.$$

Rounding the final result to three decimal places, the regression equation for predicting the language achievement measure which served as a criterion is

$$\tilde{y} = 0.475x_1 + 0.190x_2 + 0.085x_3 - 0.024x_4$$

where all scores are expressed in standard score form. Applying equation 6 of Chapter 7, the multiple correlation is found to be 0.647.



## METHOD 2—ITERATIVE SOLUTION

The basic relationship upon which the iterative method depends is the following

$$r_{ic} = \beta_1 r_{1i} + \beta_2 r_{2i} + \cdots + \beta_k r_{ki} \quad (1)$$

That is, the correlation of a test  $i$  with the criterion is equal to the sum of the products of each test's regression weight ( $\beta$ ) and its correlation with test  $i$ .

The analysis starts with the square table of obtained correlations among tests in a battery and the column of empirically determined test validities. These are given on page 336, where the problem was stated. The first step is to *guess* what the beta weight will be for each predictor variable in this set of data. Some of the considerations which enter into making a shrewd initial guess are discussed later. To the set of guessed beta weights and to the given table of intercorrelations there corresponds some set of values for the validity coefficients which will satisfy equation (1). Thus, if we designate the guessed values of the betas  $\tilde{\beta}$ , we have

$$\tilde{\beta}_1 r_{1i} + \tilde{\beta}_2 r_{2i} + \cdots + \tilde{\beta}_k r_{ki} = \tilde{r}_{ic} \quad (2)$$

The estimated set of beta weights yields a set of values for the validity coefficients, i.e., that set of validity coefficients for which this would be the exact set of beta weights.

The calculation of the  $\tilde{r}_{ic}$  values is simple and quite rapid if a calculating machine is available which will make an algebraic sum of products. The intercorrelations are set up in a square matrix with unity for the diagonal terms. A computing sheet is set up with rows spaced the same distance apart as the rows in the table of intercorrelations, and with variables numbered to correspond to the variables in the correlation table. Column I of the computing sheet contains the column of actually obtained validity coefficients. Column II contains the initial guessed beta weights. Column III contains the values of  $\tilde{r}_{ic}$ . The entries in this column are obtained by placing column II alongside each column of the correlation matrix in turn and getting the sum of products of the paired terms. The computing sheet for the illustrative example is shown in Table 2.

Once an initial set of  $\tilde{r}_c$  values has been obtained, the procedure becomes one of successive corrections to the beta weights one at a time until a set of  $\tilde{r}_c$  values is obtained which corresponds to the empirical validities,  $r_c$ , within a specified limit of accuracy. In general, one starts with the variable for which the discrepancy between  $r_c$  and  $\tilde{r}_c$  is greatest, adjusts the beta weight by an amount which will approximately eliminate the dis-

TABLE 2. COMPUTATION SHEET, ITERATIVE METHOD FOR DETERMINING REGRESSION WEIGHTS

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
Variable	$r_c$	$Wt_1$	$\tilde{r}_{c(1)}$	$\tilde{r}_{c(2)}$	$\tilde{r}_{c(3)}$	$\tilde{r}_{c(4)}$	$Wt_4$	$\tilde{r}_{c'(1)}$	$\tilde{r}_{c(5)}$	$\tilde{r}_{c(6)}$	$Wt_6$	$\tilde{r}_{c'(6)}$
1	.62	.50	.672	.642	.626	.620	.47	.621	.615	.620	.47	.620
2	.50	.20	.544	.527	.513	.503	.19	.502	.497	.502	.19	.502
3	.44	.10	.484	.467	.450	.445	.10	.449	.439	.445	.09	.445
4	.36	.00	.410	.394	.364	.359	-.03	.360	.354	.364	-.02	.364
Adjustment			Var. 1 -.03	Var. 4 -.03	Var. 2 -.01			Var. 3 -.01	Var. 4 +.01			

$$R = \frac{0.4188}{\sqrt{0.4196}} = 0.646$$

crepancy, and then computes a new set of adjusted  $\tilde{r}_c$  values. (Column IV in Table 2.) The procedure for making the adjustments is considered in the next paragraph. A second beta weight is then corrected and a new set of  $\tilde{r}_c$  values obtained, and so forth. With practice, a certain knack is developed in selecting variables to adjust and deciding upon the amount of adjustment to make. Adjustments are continued until the  $\tilde{r}_c$  and  $r_c$  values are in sufficiently close agreement. Two-place accuracy in regression weights is probably ample for any set of weights in personnel selection.

In making adjustments to the initial set of beta weights, the first principle is to adjust first the beta weight for which  $|r_c - \tilde{r}_c|$  is greatest and to adjust it by approximately the amount

$r_c - \tilde{r}_c = d_j$ . An adjustment of the weight for one variable will in general affect the values for all the  $\tilde{r}_{ic}$ 's. If we call the adjusted values  $\tilde{r}_{ic(2)}$ , we have

$$\tilde{r}_{ic(2)} = \tilde{r}_{ic} + d_j r_{ij} \quad (3)$$

This can easily be seen if we compare the terms of equation 2 for  $\tilde{r}_{ic}$  and  $\tilde{r}_{ic(2)}$ . If the adjustment  $d_j$  is a fairly small amount or a round figure such as 0.10, each correlation in column  $j$  of the correlation matrix can be multiplied by  $d_j$  mentally, the product subtracted mentally from the corresponding entry of column III of the calculation sheet, and the difference entered in column IV. Column IV then becomes the column  $\tilde{r}_{ic(2)}$  of adjusted  $\tilde{r}_c$  values. A second adjustment can be made on column IV in the same way, and so on. The beta weight to be adjusted next is always selected by comparing the column of  $\tilde{r}_c$  values resulting from the immediately preceding adjustment with the  $r_c$  column (column I) on the computing sheet and noting the location of the greatest discrepancies. A check on the accuracy of all the arithmetical processes up to that point is possible at any point by repeating the operations of formula 2 with the most recent approximation to the beta weights. (See columns VII and VIII of Table 2.)

The correlation with the criterion of a weighted score based on any specific set of weights may be computed quite simply. It is given by the formula

$$R = \frac{\sum_{i=1}^k V_i r_{ic}}{\sqrt{\sum_{i=1}^k \sum_{j=1}^k V_i V_j r_{ij}}} \quad (4)$$

where  $V_i$  signifies the weight attached to variable  $i$ . When the weights  $V_i$  correspond exactly to the regression weights, this formula simplifies to:

$$R = \sqrt{\sum \beta_i r_{ic}} \quad (5)$$

With formula 4, it is possible to determine the correlation between any set of weighted scores and an additional criterion variable. The formula can be used to find the validity of the

approximate weights at any particular stage in the approximation, as well as at the end, when the approximation has reached the desired standard of accuracy. At this point, the validity resulting from the weights will approximate very closely the multiple correlation resulting from exact regression weights. It is also possible to estimate the validity which will result from any other set of weights, if some other consideration makes it desirable to change the weighting system.

In actual computation, the numerator of formula 4 is the sum of products of the column of the latest set of weights, each times the corresponding validity coefficient in column I. The expression under the square-root sign is the sum of products of weights times corresponding  $\tilde{r}_c$  values, i.e., times the validity coefficients produced by that set of weights.

Although it would be possible, in this procedure, to start with uniform weights for all tests or with any other set of weights, a good deal of time can be saved if the initial choice is a fairly close approximation to the final weights. Kelley and Salisbury suggest starting by giving each test a weight one-half its validity coefficient. However, a little practice with the method should make possible considerably more efficient skills in that regard. Certain hints can be given to the novice with the method, as follows:

1. Initially, it often proves efficient to give a substantial number of the variables, perhaps half, zero weights. This speeds the work of preparing column III on the computation sheet.

2. If no calculating machine is available and the problem is small, initial weights should be in round numbers, i.e., 0.10, 0.20, etc.

3. The size of the initial weights depends on the number of variables. The more variables, the smaller the weights relative to the validity coefficients. With five to ten variables, the weights of those variables weighted might range from  $\frac{1}{4}$  to  $\frac{3}{4}$  of the validity coefficient. With fifteen to twenty variables, the weights might range from  $\frac{1}{4}$  to  $\frac{1}{2}$ . The highest proportions would, of course, apply to those variables with the highest validities.

4. Experience from previous work with the same criterion and/or test variables can sometimes be used as a basis for an initial set of weights.

There are also one or two tricks in applying corrections to the initial set of weights.

1. Where the bulk of the corrections required from the initial set of guessed weights are in the same direction, any given correction should be somewhat smaller than the amount  $r_{ic} - \tilde{r}_{ic}$ . This is due to the fact that when intercorrelations are largely positive the corrections on different variables tend to supplement one another.

2. Whenever corrections are being carried out by mental arithmetic, time is probably saved by making the initial large corrections in convenient amounts, such as 0.10 and 0.05.

One practical advantage of the present iterative method is that it is very simple to add any desired additional conditions to the set of weights one is deriving, and then determine the most valid set of weights which also satisfies these conditions. In much of the work in the AAF Aviation Psychology Program, the additional condition was imposed that no weights be negative. In this case, the weight for a variable would be corrected down as far as zero, but no further correction would be made. It is also a simple matter to drop out a test or group of tests (give them zero weights) and determine the weights which should then be used for the remainder of the tests. Any other desired conditions could be imposed in similar fashion.

There may be some concern that the procedures described in this section have an element of subjectivity, in that there is some choice as to which variables to adjust and how much to adjust them. However, it has been found repeatedly that different persons working with the same set of data come out with substantially the same set of weights, except for rounding errors of one point in the last place. As in the present example, the results have been found to agree, within the limits of accuracy to which the approximation is carried out, with results for the exact solution of the set of normal equations.



## APPENDIX B

### *Table for Estimating Correlations, Based on Upper and Lower 27 Per Cent of Group*

Kelley<sup>1</sup> has shown that the most accurate determination of item validities or internal consistencies can be obtained, given that the continuous nature of the criterion is to be sacrificed for computational convenience, by comparing approximately the upper and lower 27 per cent of the total group. Elimination of the middle 46 per cent leads to results that are more consistent from sample to sample than those obtained from using all available cases. On the basis of this demonstration, Flanagan<sup>2</sup> has prepared a chart, and also a table for estimating product-moment correlations from data on percentage succeeding with the task in the upper and lower 27 per cent of the group. The table, which represents the more convenient form for routine use, was privately printed by the Cooperative Test Service and is not readily available for general use. The table is therefore reproduced here (Table 3).

Use of this table is very straightforward. After the test papers or answer sheets have been scored, they are arranged in order from highest to lowest score. Twenty-seven per cent of the papers are taken, counting down from the highest score, and the same number counting up from the lowest score. (It is often convenient to work with a total group of 370, in which case 27 per cent equals 100. The computation of percentages is then facilitated.) The percentage succeeding with a given item is determined for the upper group and for the lower group, making appropriate adjustments for guessing. The table is then entered

<sup>1</sup> T. L. Kelley, "The Selection of Upper and Lower Groups for the Validation of Test Items," *J. Educ. Psychol.*, **30**, 17-24 (1939).

<sup>2</sup> J. C. Flanagan, "General Considerations in the Selection of Test Items and a Short Method of Estimating the Product-Moment Coefficient from the Tails of the Distribution," *J. Educ. Psychol.*, **30**, 674-680 (1939).

in the column corresponding to the percentage for the upper group and the row corresponding to the percentage for the lower group. The required correlation value is at the intersection of the row and column.

The table proceeds by arguments of 2 per cent in both dimensions. If the percentage succeeding on the item is odd it will, of course, be necessary to interpolate between adjacent columns or rows. If both values are odd, interpolation should be carried out between the four values which bracket the desired value. That is, if 85 per cent of the upper group and 45 per cent of the lower group succeeded with an item, the validity would be estimated to be

$$\frac{0.42 + 0.45 + 0.40 + 0.43}{4} = 0.425$$

*Table 3. A Table of the Values of the Product-Moment Coefficient of Correlation in a Normal Bivariate Population Corresponding to Given Proportions of Success*

TABLE 3. A TABLE OF THE VALUES OF THE PRODUCT-MOMENT COEFFICIENT OF CORRELATION IN A NORMAL BIVARIATE POPULATION  
(CORRESPONDING TO GIVEN PROPORTIONS OF SUCCESS)

		Proportion of successes in the 27 per cent scoring highest on the continuous variable																										
		01	02	04	06	08	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46	48	50	
01	0	11	23	30	35	40	43	46	49	51	53	55	57	59	61	62	63	65	66	67	68	69	70	71	72	72	72	
02	-11	0	12	19	25	30	34	37	40	43	46	48	50	51	53	55	56	58	59	61	62	63	64	66	67	68	68	
04	-23	-12	0	08	14	19	23	26	30	33	36	38	40	42	44	46	48	49	51	53	54	56	57	58	60	61	61	
06	-30	-19	-08	0	06	11	15	19	23	26	29	31	33	36	38	40	42	44	45	47	48	50	52	53	55	56	56	
08	-35	-25	-14	-06	0	05	09	13	17	20	23	25	28	30	32	35	37	38	40	42	44	45	47	49	51	52	48	
10	-40	-30	-19	-11	-05	0	04	08	12	15	18	21	23	26	28	30	32	34	36	38	40	41	43	45	47	48	48	
12	-43	-34	-23	-15	-09	-04	0	04	07	11	13	16	19	21	24	26	28	30	32	34	36	38	39	41	43	45	45	
14	-46	-37	-26	-19	-13	-08	-04	0	03	07	10	12	15	18	20	22	25	27	29	31	33	34	36	38	40	42	42	
16	-49	-40	-30	-23	-17	-12	-07	-03	0	03	06	09	12	14	17	19	21	24	26	28	30	31	33	35	37	39	39	
18	-51	-43	-33	-26	-20	-15	-11	-07	-03	0	03	06	08	11	13	16	18	20	23	25	27	28	30	32	34	36	36	
20	-53	-46	-36	-29	-23	-18	-13	-10	-06	-03	0	03	06	08	11	13	15	17	19	22	24	26	27	29	31	33	33	
22	-55	-48	-38	-31	-25	-21	-16	-12	-09	-06	-03	0	03	06	08	10	12	15	17	19	21	23	25	27	29	31	31	
24	-57	-50	-40	-33	-28	-23	-19	-15	-12	-08	-06	-03	0	03	05	08	10	12	14	16	18	20	22	24	26	28	28	
26	-59	-51	-42	-36	-30	-26	-21	-18	-14	-11	-08	-06	-03	0	02	05	07	09	12	14	16	18	20	22	24	26	26	
28	-61	-53	-44	-38	-32	-28	-24	-20	-17	-13	-11	-08	-05	-02	0	02	04	07	09	11	13	15	17	19	21	23	23	
30	-62	-55	-46	-40	-35	-30	-26	-22	-19	-16	-13	-10	-08	-05	-02	0	02	04	07	09	11	13	15	17	19	21	21	
32	-63	-56	-48	-42	-37	-32	-28	-25	-21	-18	-15	-12	-10	-07	-04	-02	0	02	04	07	09	11	13	15	17	19	19	
34	-65	-58	-49	-44	-38	-34	-30	-27	-24	-20	-17	-15	-12	-09	-07	-04	-02	0	02	04	06	09	11	13	15	17	17	
36	-66	-59	-51	-45	-40	-36	-32	-29	-26	-23	-19	-17	-14	-12	-09	-07	-04	-02	0	02	04	06	08	11	13	15	15	
38	-67	-61	-53	-47	-42	-38	-34	-31	-28	-25	-22	-19	-16	-14	-11	-09	-07	-04	-02	0	02	04	06	08	11	13	13	
40	-68	-62	-54	-48	-44	-40	-36	-33	-30	-27	-24	-21	-18	-16	-13	-11	-09	-06	-04	-02	0	02	04	06	08	11	10	

Proportion of successes in the 27 per cent scoring lowest on the continuous variable

42	-69	-63	-56	-50	-45	-41	-38	-34	-31	-28	-26	-23	-20	-18	-15	-13	-11	-09	-06	-04	-02	0	02	04	06	08
44	-70	-64	-57	-52	-47	-43	-39	-36	-33	-30	-27	-25	-22	-20	-17	-15	-13	-11	-08	-06	-04	-02	0	02	04	06
46	-71	-66	-58	-53	-49	-45	-41	-38	-35	-32	-29	-27	-24	-22	-19	-17	-15	-13	-11	-08	-06	-04	-02	0	02	04
48	-72	-67	-60	-55	-51	-47	-43	-40	-37	-34	-31	-29	-26	-24	-21	-19	-17	-15	-13	-11	-08	-06	-04	-02	0	02
50	-72	-68	-61	-56	-52	-48	-45	-42	-39	-36	-33	-31	-28	-26	-23	-21	-19	-17	-15	-13	-11	-08	-06	-04	-02	0
52	-73	-69	-62	-57	-53	-50	-46	-43	-40	-38	-35	-33	-30	-28	-26	-23	-21	-19	-17	-15	-13	-11	-08	-06	-04	-02
54	-74	-70	-63	-59	-55	-51	-48	-45	-42	-39	-37	-34	-32	-30	-27	-25	-23	-21	-19	-16	-14	-12	-10	-08	-06	-04
56	-75	-71	-64	-60	-56	-53	-49	-47	-44	-41	-39	-36	-34	-32	-29	-27	-25	-23	-21	-18	-16	-14	-12	-10	-08	-06
58	-76	-72	-66	-61	-58	-54	-51	-48	-45	-43	-40	-38	-36	-33	-31	-29	-27	-25	-22	-20	-18	-16	-14	-12	-10	-08
60	-77	-73	-67	-62	-59	-56	-52	-50	-47	-45	-42	-40	-37	-35	-33	-31	-29	-27	-25	-22	-21	-18	-16	-14	-12	-10
62	-78	-73	-68	-64	-60	-57	-54	-51	-49	-47	-44	-42	-39	-37	-35	-33	-31	-29	-27	-25	-22	-20	-18	-16	-15	-13
64	-78	-74	-69	-65	-61	-58	-55	-53	-50	-48	-46	-43	-41	-39	-37	-35	-33	-31	-29	-27	-25	-22	-21	-19	-17	-15
66	-79	-75	-70	-66	-63	-60	-57	-54	-52	-49	-47	-45	-43	-41	-39	-37	-35	-33	-31	-29	-27	-25	-23	-21	-19	-17
68	-80	-76	-71	-67	-64	-61	-58	-56	-53	-51	-49	-47	-45	-42	-40	-38	-37	-35	-33	-31	-29	-27	-25	-23	-21	-19
70	-81	-77	-72	-68	-65	-63	-60	-57	-55	-53	-51	-49	-46	-44	-42	-40	-38	-37	-35	-33	-31	-29	-27	-25	-23	-21
72	-82	-78	-73	-70	-66	-64	-61	-59	-57	-54	-52	-50	-48	-46	-44	-42	-40	-39	-37	-35	-33	-31	-29	-27	-26	-23
74	-82	-79	-74	-71	-68	-65	-63	-60	-58	-56	-54	-52	-50	-48	-46	-44	-42	-41	-39	-37	-35	-33	-32	-30	-28	-26
76	-83	-80	-75	-72	-69	-67	-64	-62	-60	-58	-56	-54	-52	-50	-48	-46	-45	-43	-41	-39	-37	-36	-34	-32	-30	-28
78	-83	-80	-76	-73	-70	-68	-66	-63	-61	-60	-57	-56	-54	-52	-50	-49	-47	-45	-43	-42	-40	-38	-36	-34	-33	-31
80	-84	-81	-77	-74	-72	-70	-67	-65	-63	-61	-60	-57	-56	-54	-52	-51	-49	-47	-46	-44	-42	-40	-39	-37	-35	-33
82	-85	-82	-78	-76	-73	-71	-69	-67	-65	-63	-61	-60	-58	-56	-54	-53	-51	-49	-48	-47	-45	-43	-41	-39	-38	-36
84	-86	-83	-80	-77	-75	-72	-70	-68	-67	-65	-63	-61	-60	-58	-57	-55	-53	-52	-50	-49	-47	-45	-44	-42	-40	-39
86	-87	-84	-81	-78	-76	-74	-72	-70	-68	-67	-65	-63	-62	-60	-59	-57	-56	-54	-53	-51	-50	-48	-47	-45	-43	-42
88	-87	-85	-82	-80	-77	-76	-73	-72	-70	-69	-67	-66	-64	-63	-61	-60	-58	-57	-55	-54	-52	-51	-49	-48	-46	-45
90	-88	-86	-83	-81	-79	-77	-76	-74	-72	-71	-70	-68	-67	-65	-64	-63	-61	-60	-58	-57	-56	-54	-53	-51	-50	-48
92	-89	-87	-84	-82	-81	-79	-77	-76	-75	-73	-72	-70	-69	-68	-66	-65	-63	-61	-60	-59	-58	-56	-55	-53	-52	-50
94	-90	-88	-86	-84	-82	-81	-80	-78	-77	-76	-74	-73	-72	-71	-70	-68	-67	-66	-65	-64	-62	-61	-60	-59	-57	-56
96	-91	-90	-88	-86	-84	-83	-82	-81	-80	-78	-77	-76	-75	-74	-73	-72	-71	-70	-69	-68	-67	-66	-64	-63	-62	-61
98	-92	-91	-90	-88	-87	-86	-85	-84	-83	-82	-81	-80	-80	-79	-78	-77	-76	-75	-74	-73	-73	-72	-71	-70	-69	-68
99	-93	-92	-91	-90	-89	-88	-87	-87	-86	-85	-84	-83	-83	-82	-82	-81	-80	-79	-78	-78	-77	-76	-75	-74	-73	-72

Proportion of successes in the 27 per cent scoring lowest on the continuous variable



TABLE 3. A TABLE OF THE VALUES OF THE PRODUCT-MOMENT COEFFICIENT OF CORRELATION IN A NORMAL BIVARIATE POPULATION  
(CORRESPONDING TO GIVEN PROPORTIONS OF SUCCESS (Continued))

		Proportion of successes in the 27 per cent scoring highest on the continuous variable																								
		52	54	56	58	60	62	64	66	68	70	72	74	76	78	80	82	84	86	88	90	92	94	96	98	99
01	01	73	74	75	76	77	78	78	79	80	81	82	82	83	83	84	85	86	87	87	88	89	90	91	92	93
02	02	69	70	71	72	73	73	74	75	76	77	78	79	80	80	81	82	83	84	85	86	87	88	90	91	92
04	04	62	63	64	66	67	68	69	70	71	72	73	74	75	76	77	78	80	81	82	83	84	86	88	90	91
06	06	57	59	60	61	62	64	65	66	67	68	70	71	72	73	74	76	77	78	80	81	82	84	86	88	90
08	08	53	55	56	58	59	60	61	63	64	65	66	68	69	70	72	73	75	76	77	79	81	82	84	87	89
10	10	50	51	53	54	56	57	58	60	61	63	64	65	67	68	70	71	72	74	76	77	79	81	83	86	88
12	12	40	48	49	51	52	54	55	57	58	60	61	63	64	66	67	69	70	72	73	76	77	80	82	85	87
14	14	43	45	47	48	50	51	53	54	56	57	59	60	62	63	65	67	68	70	72	74	76	78	81	84	87
16	16	40	42	44	45	47	49	50	52	53	55	57	58	60	61	63	65	67	68	70	72	75	77	80	83	86
18	18	38	39	41	43	45	47	48	49	51	53	54	56	58	60	61	63	65	67	69	71	73	76	78	82	85
20	20	35	37	39	40	42	44	46	47	49	51	52	54	56	57	60	61	63	65	67	70	72	74	77	81	84
22	22	33	34	36	38	40	42	43	45	47	49	50	52	54	56	57	60	61	63	66	68	70	73	76	80	83
24	24	30	32	34	36	37	39	41	43	45	46	48	50	52	54	56	58	60	62	64	67	69	72	75	80	83
26	26	28	30	32	33	35	37	39	41	42	44	46	48	50	52	54	56	58	60	63	65	68	71	74	79	82
28	28	26	27	29	31	33	35	37	39	40	42	44	46	48	50	52	54	57	59	61	64	66	70	73	78	82
30	30	23	25	27	29	31	33	35	37	38	40	42	44	46	49	51	53	55	57	60	63	65	68	72	77	81
32	32	21	23	25	27	29	31	33	35	37	38	40	42	45	47	49	51	53	56	58	61	64	67	71	76	80
34	34	19	21	23	25	27	29	31	33	35	37	39	41	43	45	47	49	52	54	57	60	63	66	70	75	79
36	36	17	19	21	23	25	27	29	31	33	35	37	39	41	43	46	48	50	53	55	58	61	65	69	74	78
38	38	15	16	18	20	22	25	27	29	31	33	35	37	39	42	44	47	49	51	54	57	60	64	68	73	77
40	40	12	14	16	18	21	22	25	27	29	31	33	35	37	40	42	45	47	50	52	56	59	62	67	73	77

Proportion of successes in the 27 per cent scoring lowest on the continuous variable



## Index

- Academic grades as criterion measure, 126, 154
- Adkins, D. C., 60
- Administrative problems, in conduct of testing, 257-293  
in using test results, 294-311
- Aggregate weighting board, 280-282
- Analysis of variance, related to reliability, 70-72  
technique for reliability estimate, 93-95
- Answer sheet, Clapp-Young, 275  
hand-scoring, 273-274  
IBM, 276-277, 279
- Apparatus tests, 43-44, 265-267
- Attenuation, correction for, 105
- Berkson, J., 327
- Bias in criterion measures, 130-131, 143, 147
- Bibliography of tests, 35
- Bingham, W. V., 35, 47
- Biserial correlation, 162-167, 237
- Biserial phi, 169
- Brinton, W. C., 330
- Buros, O. K., 35
- Canonical correlation, 211
- Cattell, J. McK., 36
- Chesire, L., 168
- Clapp-Young answer sheet, 275
- Classification of personnel, basis for administrative decision in, 302-305  
use of tests for, 8  
problem defined, 221-222  
qualities desired, 224-226  
with two tests, 222-224
- Clinical method applied to test scores, 200-201
- Composite score, calculation of, 280-283  
checking of, 284-285  
types of, 307-310
- Correction for attenuation, 105
- Correlation, biserial, assumptions, 163  
computation of, 162  
in item analysis, 237-238  
when to use, 164-166, 167  
canonical, 211  
from upper and lower 27 per cent, 242, 345-350  
multiple, computation procedures, 335-344  
factors determining, 190-193  
formula, 188, 342  
shrinkage formula, 204  
suppression variable in, 192-193  
point biserial, computation of, 164  
in item analysis, 237-238  
when to use, 164-166, 167  
product-moment, 161  
tetrachoric, 168, 240
- Correlation ratio, 181
- Cost-utility, 327
- Criterion, academic grades as, 154-155  
administrative action as, 157-159  
bias as a factor in, 130-131, 143, 147  
dichotomous, 161  
importance of, 119  
objective performance score as, 137-141

Criterion (*Continued*)

- objectivity of measures, 133-136
  - observer-scored job sample as, 141-145
  - partial, combining, 210-212
  - performance check as, 137-141
  - practicality as factor in, 131
  - qualities desired in, 124-132
  - rating, of job sample, 145-149
    - summary, 155-157
  - rational vs. empirical considerations, 123-124
  - relevance as a factor in, 125-127
  - reliability, as a factor in, 106-108, 127-131
    - of different types, 141, 144, 148, 153
  - specific test as, 132, 133-149
  - summary evaluation as, 132, 149-159
  - summary performance records as, 152-154
  - test of knowledge as, 133, 136-137
  - types of, 132-159
  - ultimate vs. immediate, 121-124
- Critical requirements, 15
- Curtailment, effect on validity, 170-171
  - generalized correction formula, 176
  - need to correct for, 168-170
  - Pearson correction formulas, 172-175
  - with dichotomous criterion, 177-180
- Cutoff, multiple, as technique of combining tests, 195-200
- Cutting score in multiple prediction, 195-200
- Davis, F. B., 41, 98, 236, 242
- Dichotomy, as criterion, 161
  - correlation between two, 167-169
  - correlation with, 162-167
  - curtailment with, 177-180

- Difficulty, *see* Item difficulty
- Doolittle solution, abbreviated procedure, 336-339
- Dwyer, P. S., 335
- Eisenhart, C., 203
- Face validity, 4, 33
- Factor analysis, in job analysis, 30
  - in test construction, 219-220
- Fear, R. A., 47
- Ferguson, G. A., 92
- Fisher, R. A., 181
- Flanagan, J. C., 242, 249, 345
- Franzen, R., 197
- Garrett, H. E., 34
- Gerberich, J. R., 34
- Gibson, J. J., 42
- Gillman, L., 177
- Goode, H. H., 177
- Greene, E. B., 34
- Greene, H. A., 34
- Guilford, J. P., 30, 41, 220
- Hawkes, H. E., 60
- Hildreth, G., 35
- Horst, A. P., 176, 207, 250
- Hotelling, H., 211
- Hoyt, C., 93
- Internal consistency, situations for which appropriate, 230-231
  - use of data, 252-256
- International Business Machines, aggregate weighting board, 280-282
  - tabulating equipment, 283
  - test-scoring machine, 273, 276-279
- Interview, in job analysis, 25-27
  - in personnel selection, 46-48
- Item analysis, 5, 53, 227-256
- Item difficulty, correction formula, 234

- Item difficulty (*Continued*)  
 indices of, 233-236  
 optimum, 229  
 significance of, 228-230
- Item discrimination, indices of, with  
 continuous score, 237-239  
 with dichotomous groups, 239-240  
 with extreme groups, 240-243  
 internal consistency vs. validity, 230-232
- Item validity, type of test for which  
 needed, 230-231  
 ways of using, 243-252
- Item writing, 60-67, 254-256
- Iterative solution for regression  
 equation, 340-344
- Jackson, R. B. W., 92
- Job analysis, 3, 12-31  
 categories for, 16-18  
 factor analysis in, 30  
 in selecting criterion measures, 12  
 procedures for collecting information, analysis of documentary material, 20-25  
 analysis of test validities, 29-31  
 direct experience, 27-29  
 interviews and interrogations, 25-27  
 use of previous studies, 18-20  
 purpose of, 13
- Job description, 14-16
- Jordan, B., 47
- Jorgensen, A. N., 34
- Kelley, T. L., 188, 241, 336, 345
- Kuder, G. F., 91
- Lindquist, E. F., 60
- Mann, C. R., 60
- Mental Measurements Yearbook*, 35
- Moore, B. V., 47
- Moore, H., 35
- Motion picture tests, 41-42, 267-268
- Motivation as problem in testing, 57, 270-272
- Multiple correlation, *see* Correlation, multiple
- Multiple cutoff, 195-200
- Non-linear function in multiple prediction, 194-195
- Non-linear relationship, 180-184
- Non-test variables, use in personnel selection, 305-307
- Observation as selection technique, 47-48
- Office of Strategic Services, 48
- Paterson, D. G., 34
- Pearson, K., 172
- Performance test as criterion measure, 137-145
- Personnel procedures, daily quota approach, 299-305  
 predicted yield approach, 299-305  
 showing effectiveness of, 312-334  
 supply and demand as factor, 296-299  
 time as factor, 296-299
- Personnel psychologist, administrative functions of, 310-311  
 and the employee, 333-334  
 as salesman, 312-334
- Personnel selection, *see* Selection of personnel
- Peters, C. C., 182
- Phi coefficient, 169, 240
- Point biserial correlation, 164-167, 237-238
- Predicted yield, 299-305
- Printed tests, administration, 261-265  
 advantages of, 39-41  
 as criterion, 132



- Proficiency records, as criterion,  
152-154  
use in job analysis, 21-25
- Psychological Corporation, 36
- Public relations, 312-334
- Range, restriction of, *see* Curtailment
- Rating, as criterion measure, 145-149, 155-157  
as selection procedure, 44-49  
reliability of, 116-118
- Record systems, card layout, 290-292  
filing system, 287-288  
functions of, 286-287  
type of score in, 292  
use of IBM equipment, 288-290
- Reeve, E., 176
- Regression, multiple, equation, 185-187  
use in selection, 185-186
- Regression weights, 189-190  
Doolittle solution, 188, 336-339  
iterative solution, 188, 340-344
- Reliability, ability level as a factor, 98  
absolute, 69, 102-104  
alternate trials vs. alternate periods, 115  
analysis of variance, and logic of, 70-72  
estimate of, 90-96  
coefficient of, 69, 71, 104  
correction for length of test, 84  
factors influencing, 96-102  
from equivalent tests, 79-81  
heterogeneity as a factor, 96-98  
Hoyt's procedure, 93-95  
independence in ratings, 116-118  
interpretation of estimates, 102-104  
item characteristics as a factor, 99-101
- Reliability (*Continued*)  
Kuder-Richardson formulas, 91-93  
learning tasks, 112-114  
length of test as a factor, 84, 101-102  
logical considerations in evaluating, 69-78  
procedures for estimating, 78-96  
relative, 69, 104  
retest, 81-83  
special problems, 111-118  
speeded tests, 112  
split-test, 82-99  
alternate item groups, 89  
alternate items, 89  
first vs. second half, 90  
rationally equivalent halves, 87-88  
separately timed halves, 88  
use of data, 104-111  
in analyzing tests, 108-111  
in evaluating criterion, 106-108, 127-131  
validity, effect upon, 109  
when result is known, 115
- Restriction of range, *see* Curtailment
- Richardson, M. W., 91
- Rider, P. R., 203
- Russell, J. R., 325
- Saffir, M., 168
- Salisbury, F. S., 188, 336
- Schneck, M. R., 34
- Schneider, G. G., 34
- Score, types of, 292, 307-310
- Security of test materials, 33, 268-270
- Selection of personnel, multiple, 217-220, 295  
rationale of, 214-217  
techniques for, clinical, 200-201  
linear regression, 185-194  
multiple cutoff, 195-200  
non-linear function, 194-195

Selection of personnel (*Continued*)  
 use of non-test data in, 305-307  
 vs. classification, 294-296  
 Shartle, C. L., 13, 19, 202, 204  
 Shrinkage in multiple correlation,  
 204  
 Standard error of measurement, 69,  
 102-104  
 Stead, W. H., 202, 204  
 Supply and demand as factors in  
 personnel procedures, 296-  
 299  
 Suppression variable, 192-193  
 Symonds, P. M., 34  
 Taylor, H. C., 325  
 Test, apparatus, problems in using,  
 265-267  
 values of, 43-44  
 battery, addition to, 202, 205-  
 210  
 job approach to assembling, 37  
 length of, 201-205  
 trait approach to assembling,  
 37  
 use for selection, 185-201  
 bibliography, 35  
 choice of, 4  
 combination into a battery, 7  
 equivalent forms of, 79-81  
 factor analysis in, 219-220  
 format, 66-67  
 invention of, 4, 50  
 items, editing of, 254-256  
 types of, 60-63  
 writing of, 63-66  
 media for, 39-44  
 apparatus, 43-44  
 motion picture, 41-42  
 printed, 39-41  
 motion picture, problems in using,  
 267-268  
 values of, 41-42  
 preliminary tryout of, 5, 51-54  
 printed, administration of, 261-  
 265

Test, printed (*Continued*)  
 advantages of, 39-41  
 pure vs. complex, 37, 216, 218-  
 219, 225-226  
 reviews of, 35  
 revision of, 53-54  
 scoring procedures, checking, 284-  
 285  
 hand, 273-276  
 machine, 276-279  
 sources of information about, 33-  
 36  
 specifications for, 50-51, 80  
 speed, reliability of, 112  
 uniqueness as factor in, 110, 206  
 Testing, administrative problems in,  
 257-293  
 motivation of subjects in, 270-272  
 program, promotion of, 312-314  
 types of, 213-216  
 records from, 286-292  
 schedule for, 259-261  
 sequence of tests, 261  
 time allotment, 259  
 use of results from, 294-311  
 Tetrachoric correlation, 168, 240  
 Thomson, G. H., 220  
 Thurstone, L. L., 17, 168, 220  
 Time as factor in personnel proce-  
 dures, 296-299  
 Tyler, R. W., 60  
 Uniqueness of test, 110, 206  
 Validation testing, 5-7, 54-60  
 choice of groups for, 54-58  
 numbers required for, 58-60  
 optimum time for, 54-58  
 Validity, additional contributed by  
 a test, 202-204, 207-208  
 differential, 221-226  
 indices of, analytical, 160-169  
 practical, 323-330  
 item, 230-231, 243-252

Validity (*Continued*)

- representation of, 314-330
- use of data in job analysis, 29-31
- Van Voorhis, W. R., 182
- Variance, analysis of, *see* Analysis of variance
- sources of, in test scores, 72-78
- Wang, C. K. A., 35
- Weighted score, *see* Composite score
- Weights, regression, *see* Regression weights
- Wherry, R. J., 202, 204
- Whipple, G. M., 34
- Williamson, E. G., 34